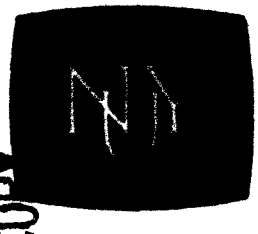
No. 196047

IMM-NYU 266
APRIL 1960



NEW YORK UNIVERSITY
INSTITUTE OF
MATHEMATICAL SCIENCES

A UNIFIED THEORY OF ESTIMATION. I.

(Revised and extended, February 1960)

ALLAN BIRNBAUM

PREPARED UNDER
CONTRACT NO. NONR-285(38)
WITH THE
OFFICE OF NAVAL RESEARCH
UNITED STATES NAVY

IMM-NYU 266
April 1960

A UNIFIED THEORY OF ESTIMATION. I.

(Revised and extended, February 1960)

Allan Birnbaum

This report represents results obtained at the Institute of Mathematical Sciences, New York University, under the sponsorship of the Office of Naval Research, Contract No. Nonr-285(38). Some sections include results previously reported under the same title, obtained at Columbia University under the sponsorship of the Office of Naval Research, Contract No. Nonr-266(33).

CONTENTS

0. Introduction and Summary. This paper extends and unifies some previous formulations and theories of estimation for one-parameter problems. The basic criterion used is admissibility of a point estimator, defined with reference to its full distribution rather than special loss functions such as squared error. Theoretical methods of characterizing admissible estimators are given, and practical computational methods for their use are illustrated in a variety of examples.

Point, confidence limit, and confidence interval estimation are included in a single theoretical formulation, and incorporated into estimators of an "omnibus" form called "confidence curves." The usefulness of the latter for some applications as well as theoretical purposes is illustrated.

Fisher's maximum likelihood principle of estimation is generalized, given exact (non-asymptotic) justification, and unified with the theory of tests and confidence regions of Neyman and Pearson. Relations between exact and asymptotic results are discussed.

An application of the general theory gives optimal sequential estimators having prescribed precision in a specified interval.

Further developments, including multiparameter and nuisance parameter problems, problems of choice among admissible estimators, formal and informal criteria for optimality, and related problems in the foundations of statistical inference, will be presented subsequently.

1. A broad formulation of the problem of point estimation. We consider problems of estimation with reference to a specified experiment E, leaving aside here questions of experimental design including those of choice of a sample size or a sequential sampling rule; some definite sampling rule, possibly sequential, is assumed specified as part of E. Let $S = \{x\}$ denote the sample space of possible outcomes x of the experiment. Let $f(x,\theta)$ denote one of the elementary probability functions on S which are specified as possibly true. Let $\Omega = \{\theta\}$ denote the specified parameter space. For each θ in Ω and for each subset of A of S, the probability that E yields an outcome x in A is given by

$$\text{Prob}\{X \in A | \theta\} = \int_A f(x,\theta)\, d\mu(x),$$

where μ is a specified σ- finite measure on S. (We assume tacitly here and below that consideration is appropriately restricted to measurable sets and functions only.)

If $\gamma = \gamma(\theta)$ is any function defined on Ω (e.g. $\gamma(\theta) \equiv \theta$ or $\gamma(\theta) \equiv \theta^2$), with range Γ, a point estimator of γ is any measurable function $g = g(x)$ taking values in Γ (or in $\bar{\Gamma}$, its closure, if, for example, Γ is an open interval). The problem of choosing a good estimator, that is an estimator which tends to take values close to the true unknown value of γ, has been formulated mathematically in various ways. Most formulations achieve mathematical definiteness by introducing criteria of closeness which appear somewhat arbitrary from some standpoints of application and undesirably schematic as expressions of the intuitive notion of closeness.

If Ω is given no specific (parametric) structure, then the latter features can be fully avoided only by a very broad formulation

which specifies only that if γ is true, then an exactly correct estimate ($g = \gamma$) is closer than any incorrect estimate ($g \neq \gamma$). If Ω is finite, $\Omega = \omega_1, \ldots \omega_k$, and $\gamma(\omega) = \omega$, this leads to the formulation of Lindley [1] in which estimators are compared only on the basis of their error probabilities

$$p_{ij} = \text{Prob}\{\omega^*(X) = \omega_i | \omega_j\}, \quad i,j, = 1,\ldots k, \quad i \neq j,$$

where $\omega^*(x)$ is any estimator of ω. This formulation has no very useful extension to typical estimation problems in which, for example, Ω is an interval, and in which the event $\omega^*(X) = \omega$ exactly has typically negligible probability and little interest.

The case in which Ω is any set of real numbers, for example an interval, and $\gamma(\omega) \equiv \omega$, may be termed the central problem of theory of point-estimation, although very important generalizations of this problem have been treated extensively. For this problem, closeness of ω^* to ω has been specified by the introduction of specific loss functions: The absolute error criterion, $|\omega^* - \omega|$, was introduced by Laplace. Gauss replaced this by the squared error criterion $(\omega^* - \omega)^2$ which proved mathematically much more tractable and provided a definite formulation of the problem which seemed equally reasonable. A generalized squared error criterion, $c(\omega) \cdot (\omega^* - \omega)^2$, where $c(\omega)$ is any specified positive function, is used in some work in modern statistical decision theory. Such criteria are sometimes used in conjunction with the requirement of unbiasedness, $E(\omega^*(X)|\omega) \equiv \omega$; this is done (evidently primarily to facilitate mathematical developments) particularly in the theory of linear estimation due to Gauss; this reduces the mean squared

error criterion to a criterion of variance: $E[(\theta^*-\theta)^2|\theta] \equiv \mathrm{Var}(\theta^*|\theta)$. (For a brief account of the history of the theory of point estimation, cf. Neyman [2], pp. 9-14.)

Each such definite specification of closeness can be criticized as somewhat arbitrary, except in a context where one postulates the reality of the indicated costs of errors of each possible kind. To avoid such features it is evidently necessary and sufficient to adopt the following weak specification of closeness: If $\theta_1^* < \theta_2^* \leq \theta$ or if $\theta \leq \theta_2^* < \theta_1^*$, the estimate θ_2^* is called closer than θ_1^* to θ; if $\theta_1^* < \theta < \theta_2^*$, no comparison as to closeness is to be made. (The latter point was put forth by Galileo in an exchange which retains interest in connection with questions of formulation of estimation problems, particularly distinctions between errors of inference and economic valuations, and the historical origins of unbiasedness criteria. Cf. [3].)

This specification of closeness leads to comparisons between estimators on the basis of all of their probabilities of errors of over-estimation and under-estimation by various amounts $d = |\theta^*-\theta|$:

$$a(u,\theta,\theta^*) = \begin{cases} F(u,\theta,\theta^*) \equiv \mathrm{Prob}\{\theta^*(X) \leq u|\theta\} & \text{for } u < \theta, \\ 1-F(u-\theta,\theta,\theta^*) \equiv \mathrm{Prob}\{\theta^*(X) \geq u|\theta\} & \text{for } u > \theta. \end{cases}$$

That is, estimators are compared only on the basis of their complete cumulative distribution functions (c.d.f's.) $F(u,\theta,\theta^*)$ for each $\theta \in \Omega$, rather than on the basis of certain "summaries" (functionals) of these c.d.f's such as mean squared error. The function $a(u,\theta,\theta^*)$, defined for any estimator $\theta^*(x)$ at each $\theta \in \Omega$ and each $u \neq \theta$, will be called the <u>risk curve</u> of θ^* at θ (or, more precisely, of $\theta^*(.)$ at θ).

The family of distributions under consideration may be viewed as having a parametric structure only in the sense that it is ordered by the labeling of each function $f(x,\Theta)$ of x by a different real number Θ. From this standpoint, the problem of estimating Θ is equivalent to that of estimating $\gamma = \gamma(\Theta)$ if the latter is any specified strictly monotone function. The formulation adopted above is clearly unaffected by (invariant under) such transformations of the parameter space ($\Omega \to \gamma(\Omega) \equiv \Gamma$), as contrasted with some other formulations referred to above.

A theory of point estimation based on this broad formulation seems appropriate for typical problems of inference occurring in empirical research, since various kinds of errors of inference and their probabilities admit simple direct interpretations, whereas other formulations introduce specifications akin to costs of various errors which seem somewhat hypothetical or arbitrary in such situations. The present theory also has theoretical and technical relevance for estimation theories based on more restrictive formulations, since it includes such theories in a formal sense which will be elaborated in a following section.

2. Admissible point estimators. An estimator $\Theta^*(x)$ of Θ is naturally considered a good one if its error-probabilities are suitably small, i.e. if (the ordinates of) its risk curves $a(u,\Theta,\Theta^*)$, for each $\Theta \in \Omega$ and each $u \neq \Theta$, are suitably small. This leads to a natural partial ordering of estimators, under which some but not all pairs of estimators can be compared. As a basis for systematic evaluations and comparisons of estimators we require the following

Definitions: For a given estimation problem, an estimator Θ^* is called at least as good as an estimator Θ^{**} if $a(u,\theta,\Theta^*) \leq a(u,\theta,\Theta^{**})$ for all $\theta \in \Omega$ and all $u \neq \theta$. If Θ^* and Θ^{**} are each each at least as the other, then $a(u,\theta,\Theta^*) \equiv a(u,\theta,\Theta^{**})$, and the estimators are called equivalent. If neither of Θ^*, Θ^{**} is at least as good as the other, the two estimators are called not comparable. If Θ^* is at least as good as Θ^{**} and if $a(u,\theta,\Theta^*) < a(u,\theta,\Theta^{**})$ for some $\theta \in \Omega$ and some $u \neq \theta$, Θ^* is called better than Θ^{**}. As estimator Θ^* is called admissible if no other estimator is better than Θ^*. The class of admissible estimators is called the admissible class. A class of estimators is called complete if, for each estimator outside the class, there is a better one in the class. The minimal (smallest) complete class, if one exists, coincides with the admissible class. A class of estimators is called essentially complete if, for each estimator not in the class, there is one at least as good in the class. A minimal essentially complete class, if one exists, is a subclass of the admissible class.

The above definition of admissibility was included in a list of criteria for point estimators by Savage [4] (pp.224-225), but it has not previously been used systematically.

The criterion of closeness of estimators introduced by Pitman [5] also deals with the full c.d.f's. of estimators, in the form of the joint distribution of each pair of estimators being compared; however this criterion does not give a partial ordering of estimators, and does not lend itself to our present purposes.

For the probabilities of under-estimation and over-estimation, we define also

$$a(\theta-,\theta,\theta^*) = \text{Prob}\left\{\theta^*(X) < \theta|\theta\right\} = \lim_{\substack{\varepsilon \to 0,\\ \varepsilon > 0}} a(\theta-\varepsilon;\theta,\theta^*),$$

$$a(\theta+,\theta,\theta^*) = \text{Prob}\left\{\theta^*(X) > \theta|\theta\right\} = \lim_{\substack{\varepsilon \to 0,\\ \varepsilon > 0}} a(\theta+\varepsilon;\theta,\theta^*).$$

For formal convenience, we also define $a(\theta,\theta,\theta^*) \equiv 0$. When reference to a given estimator θ^* is understood, we may write simply $a(u,\theta)$, $a(\theta-,\theta)$, or $a(\theta+,\theta)$. The functions $a(\theta-,\theta)$ and $a(\theta+,\theta)$ of θ play a useful technical role, and will be called respectively the lower and upper location functions of θ^*.

In many problems, estimators for which $\text{Prob}\left\{\theta^*(X) = \theta|\theta\right\} > 0$ for some θ are found not useful. The remaining estimators have continuous c.d.f's,, and have $a(\theta-,\theta) \equiv 1-a(\theta+,\theta)$. No two such estimators, having different location functions, can be comparable; for $a(\theta-,\theta,\theta^*) < a(\theta-,\theta,\theta^{**})$ is equivalent to $a(\theta+,\theta,\theta^*) > a(\theta+,\theta,\theta^{**}$ this shows that neither estimator is at least as good as the other.

The broad and "weak" definition of admissibility adopted here leads to very large admissible classes in typical problems. However it does not seem unreasonable to conceive of the problem of point estimation as one in which the investigator chooses an estimator on the basis of consideration of the risk curves of all estimators in some essentially complete class. In principle this consideration should be complete, but of course the practical counterpart of this can be at most a more or less extensive familiarity with an essentially complete class, developed by study of the risk-curves of a variety of specific estimators, possibly strengthened by some general theoretical considerations (including envelope risk-curves, discussed below), and perhaps also by reference to one or several loss

functions and criteria of optimality which may seem more or less appropriate in specific applications. Such an approach is not so difficult to carry out as might be anticipated, as will be illustrated. Of course difficulties of computation or complexity may sometimes dictate that an inadmissable estimator must be adopted; even in such cases, the most general basis on which any particular estimator might be justified as not too inefficient, is evidently the comparison of its risk-curves with those of other estimators, especially admissible ones.

Example. Let X be normally distributed with unknown mean θ and variance 1, with $\Omega = \{\theta \mid -\infty < \theta < \infty\}$. Consider, when $\theta = 1$, the risk curves of the classical estimator $\hat{\theta}(x) = x$, and of the estimators $\theta^*(x) = x + 1$ and $\theta^{**}(x) \equiv +\infty$. We have

$$a(u,1,\theta) = \begin{cases} \Phi(u-1) & \text{for } u < 1, \text{ and} \\ 1 - \Phi(u-1) & \text{for } u > 1, \end{cases}$$

where

$$\Phi(v) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{v} e^{-\frac{v^2}{2}}\, dv,$$

$$a(u,1,\theta^*) = \begin{cases} \Phi(u-2) & \text{for } u < 1, \\ 1 - \Phi(u-2) & \text{for } u > 1, \end{cases}$$

and

$$a(u,1,\theta^{**}) = \begin{cases} 0 & \text{for } u < 1, \\ 1 & \text{for } u > 1. \end{cases}$$

Our wishful goal in choosing an estimator would be to minimize simultaneously all ordinates of such curves, for all θ and all $u \neq \theta$, since each ordinate is the probability of an error. Of course this goal cannot be realized in non-trivial problems. The

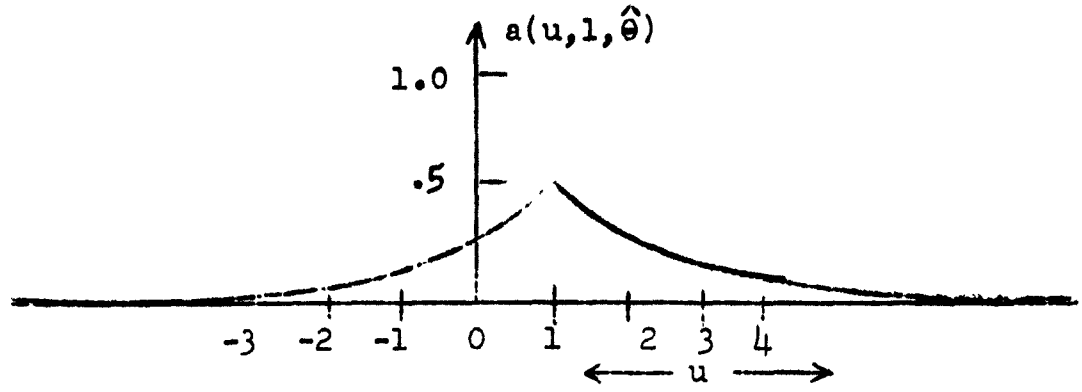
estimator θ^* is superior to $\hat{\theta}$ with respect to all errors of under-estimation, but worse with respect to over-estimation. From this standpoint neither can be called better than the other; they are not comparable. The apparently trivial estimator θ^{**} (but no "smaller" one) is perfect in avoiding errors of under-estimation, but is as bad as possible with respect to over-estimation.

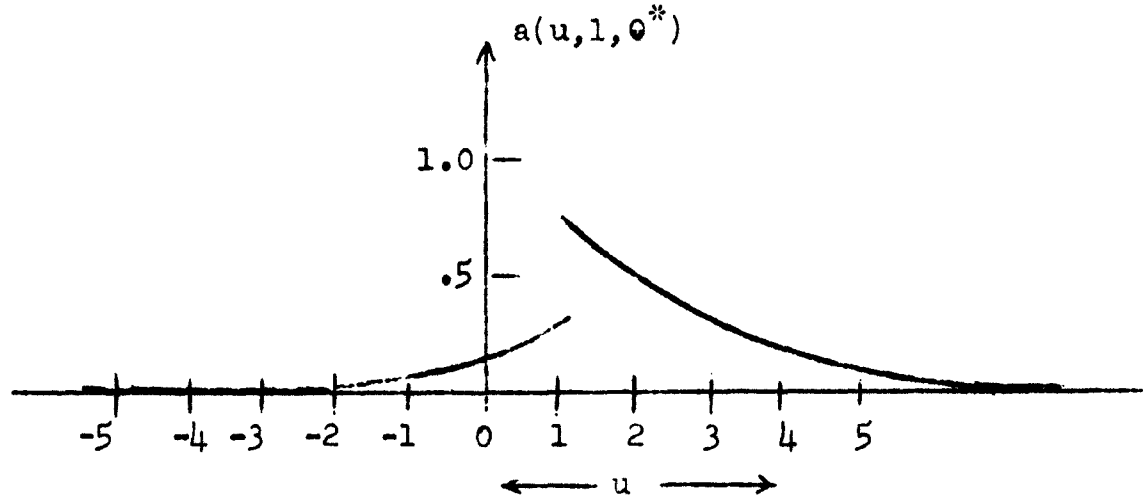
It will be seen below that each of these estimators is not only admissible but that each has, among all estimators with the same location functions, uniformly smallest risk curves.

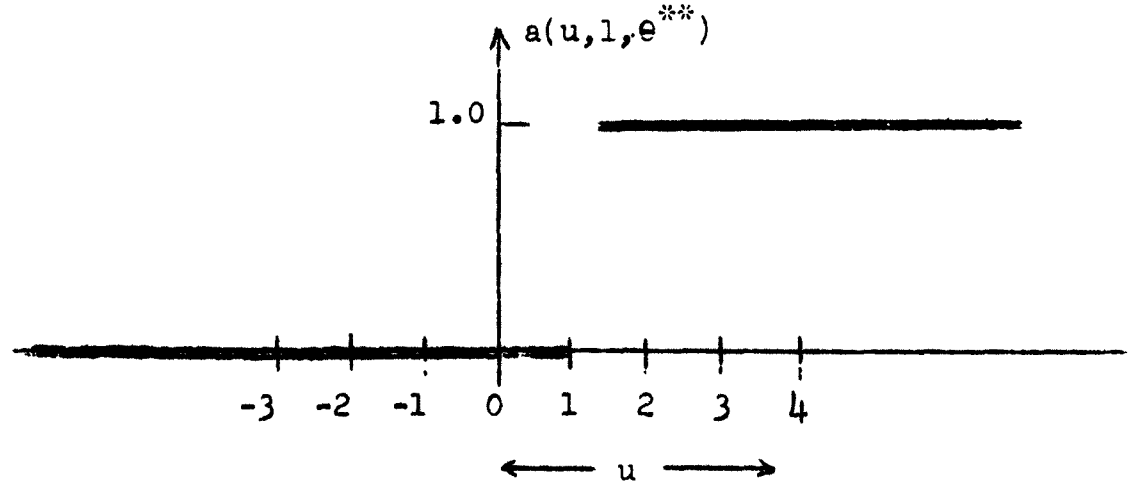
In most decision-theoretic formulations of statistical problems a real-valued risk function $r(\theta,\theta^*)$ is defined for each parameter point and each decision function. In the present formulation, we associate with each pair θ, θ^* a <u>set</u> of error-probabilities $a(u,\theta,\theta^*)$, $u \neq \theta$. These respective error-probabilities, for each fixed θ and θ^*, may be regarded as components of a vector denoted by $r(\theta,\theta^*) = \{a(u,\theta,\theta^*)\}$, the components $a(u,\theta,\theta^*)$ having index u. Then $r(\theta,\theta^*)$ is an example of a vector-valued risk function.

Knowledge of the admissible class or of an essentially complete class of estimators in the present broad sense can be useful in applying other formulations of the estimation problem. For example, every estimator which is admissible with respect to a squared error loss function must clearly be admissible in the present sense; hence the search for estimators good in the former sense can be restricted without loss to any class known to be essentially complete in the broader sense. In this way, a hierarchy of definitions of admissibility leads to a corresponding nested hierarchy of admissible or essentially complete classes of estimators. (The latter concepts,

Figure 1


a(u,1,θ̂)
1.0
.5
-3 -2 -1 0 1 2 3 4
u


a(u,1,θ*)
1.0
.5
-5 -4 -3 -2 -1 0 1 2 3 4 5
u


a(u,1,θ**)
1.0
-3 -2 -1 0 1 2 3 4
u

and that of vector-valued risk functions, were introduced in other contexts by L. Weiss [6].)

3. Admissible confidence limits. If $\Theta'' = \Theta''(x)$ is a point estimator of Θ in a specified problem, with the property that $\text{Prob}\,[\Theta''(X) < \Theta \mid \Theta] = a(\Theta-,\Theta,\Theta'')$ is relatively small for all Θ, then Θ'' is an upper estimator of Θ. In particular, if $a(\Theta-,\Theta,\Theta'')=\alpha$ for all Θ, then Θ'' is an upper confidence limit with confidence coefficient $1 - \alpha$, or an upper $(1-\alpha)$ confidence limit. Typically a value $(1-\alpha)>>.5$ is chosen.

The typical use and interpretation of an upper estimate is the following: When a given numerical value (observed value) is obtained by use of an upper estimator, this is taken as evidence supporting the conclusion or decision that the true unknown value is at least as small as the estimated value. Hence the merits of any upper estimator depend upon the following considerations, in suitable combination:

(a) The probability should be suitably high that the indicated conclusions, of the form: "Θ is not greater than $\Theta''(x)$," are correct for each possible true value of Θ. That is, the confidence coefficient should have a suitably large value; or, more generally, the lower location function $a(\Theta-,\Theta,\Theta'')$ should have suitably low values for all Θ. Such properties are sometimes referred to by the term validity, particularly in the case of confidence limit estimators; a valid $(1-\alpha)$ upper confidence limit estimator is one which does in fact have the property that $\text{Prob}\left\{\Theta^* < \Theta \mid \Theta\right\} = \alpha$ for all $\Theta \in \Omega$.

(b) Given that one of the indicated conclusions ("$\Theta \leqq \Theta''(x)$") is correct, it should be as strong and informative a conclusion as possible; hence for each possible true value of Θ, the conditional

distribution of $\Theta''(X)$, given that $\Theta \leqq \Theta''(X)$, should be concentrated as close to Θ as possible. That is, given the location function $a(\Theta-,\Theta,\Theta'')$ of any upper estimator Θ'', for each Θ and each $u > \Theta$ the values $a(u,\Theta,\Theta'') = \text{Prob}\,[\Theta''(X) \geqq u|\Theta]$ should be suitably small. Such properties of confidence limits have been termed <u>accuracy</u> properties by Lehmann [7], p.78. More generally, in the theory of confidence region estimation, such properties have been termed <u>shortness</u> properties by Neyman [8].

(c) Given that one of the indicated conclusions ("$\Theta \leqq \Theta''(x)$") is incorrect (i.e. that in fact $\Theta > \Theta''(x)$), the indicated conclusion should be misleading in the smallest possible degree. For example, in any given problem, under any given true value of Θ, when an upper estimator takes a value two units below the true value, the indicated conclusions (or inferences or actions or decisions) are at least as erroneous (or inappropriate) and in general more so, than when an upper estimator (with the same confidence coefficient or location function) takes a value which is only one unit below the true value. That is, given the location function $a(\Theta-,\Theta,\Theta'')$, for each Θ and each $u < \Theta$ the values $a(u,\Theta,\Theta'')$ should be suitably small. This property has evidently not previously been discussed along with those of validity and shortness, but it seems necessary to include it for a complete specification of the practical purposes and intuitive goals of confidence limit estimation. All three properties are given some weight in a specific loss function adopted in the decision-theoretic treatment of Wolfowitz [9].

These considerations lead in the usual way to definitions of admissibility and of complete classes of upper and lower estimators. Properties (b) and (c) together are formally identical with the

closeness properties considered in the preceding section for point estimators, while property (a) by itself is merely descriptive of the location function of a point estimator. Thus every admissible confidence limit estimator is, formally, an admissible point estimator as defined above, and is contained in every complete class of point estimators.

Hence there is no necessary formal distinction between the formulations, theories, and practical techniques of point estimation on the one hand and of confidence limit estimation on the other: the distinctions required here are only those of qualitative emphasis and quantitative degree which reflect the variety of possible purposes for which a point or confidence limit estimator may be chosen from, say, an essentially complete class. For example, in choosing an upper estimator for a given application, it may be judged that property (c) above should be given no weight as compared with properties (a) and (b) because "a miss is as good as a mile" in the given context of application; in other contexts, including probably most cases of estimation for informative inference, some weight may be given to each property.

4. Admissible interval estimators. If $J = J(x) = (\Theta', \Theta'') = (\Theta'(x), \Theta''(x))$ is a pair of point estimators such that $\Theta'(x) \leq \Theta''(x)$ for each x in S, then J is an interval estimator of Θ. In partiuclar, if $\text{Prob}\left\{\Theta'(X) \leq \Theta \leq \Theta''(X) \mid \Theta\right\} = 1-\alpha$ for each Θ, then J is a confidence interval with confidence coefficient $1-\alpha$, or a $(1-\alpha)$ confidence interval. (Typically a value $(1-\alpha) >> .5$ is chosen.) The typical use and interpretation of an upper estimate is the following: When given numerical values Θ' and Θ'' are obtained by use of an

interval estimator, this is taken as evidence for the conclusion that the true unknown value of the parameter θ lies in the closed interval $[\theta', \theta'']$.

The probability properties of any interval estimator J may be described in the following terms: It is natural to call $a(\theta-, \theta, \theta'')$ the lower location function of J (as well as of θ''), and to denote it when convenient by $a(\theta-, \theta, J)$; similarly $a(\theta+, \theta, J) \equiv a(\theta+, \theta, \theta')$ is the upper location function of J. As with point estimators, these functions give respectively the probabilities of underestimation and of overestimation when a given interval estimator J is used. For example, it is natural, to call J a <u>median-unbiased</u> interval estimator if for each θ we have equal probabilities of overestimation and underestimation: $a(\theta-, \theta, J) = a(\theta+, \theta, J)$. This usage is compatible with the definition of a median-unbiased point estimator.

A quantity of primary interest is the probability that the conclusion indicated by any interval estimator J ("θ lies in $[\theta', \theta'']$") will be incorrect, for each possible true value θ. This probability is just the sum of the location functions of J:

$$\text{Prob}\left\{\theta \text{ not covered by } J(X) \mid \theta\right\} = \text{Prob}\left\{\theta''(X) < \theta \mid \theta\right\}$$
$$+ \text{Prob}\left\{\theta(X) > \theta \mid \theta\right\} = a(\theta-, \theta, J) + a(\theta+, \theta, J).$$

If this probability equals α for each θ, then J is a $(1-\alpha)$ confidence interval; if in addition J is median-unbiased, then θ' and θ'' are $(1-\frac{1}{2}\alpha)$ confidence limits. As with point and confidence limit estimators, it is of interest in general to consider the probabilities of errors of under-estimation and of over-estimation of various magnitudes in interval estimation; we denote these probabilities by

$$a(u,\theta,J) = \begin{cases} a(u,\theta,\theta') \text{ for each } u > \theta, \\ a(u,\theta,\theta'') \text{ for each } u < \theta. \end{cases}$$

In a formal sense, a point estimator may be regarded as an interval estimator $J = (\theta', \theta'')$ having the special form: $\theta'(x) = \theta''(x)$ for all x. The full specification of what is meant by a good point estimator θ^*, by use of the risk curves $a(u,\theta,\theta^*)$, corresponds to the use of the functions $a(u,\theta,J)$ to specify at least part of what is meant by a good interval estimator J.

Again, in a formal sense an upper estimator $\theta''(x)$ may be regarded as an interval estimator $J = (\theta',\theta'')$ having the special form: $\theta'(x) \equiv \underline{\theta}$ = the greatest lower bound of Ω, for all x. The full specification of what is meant by a good upper estimator θ'', by use of the risk curves $a(u,\theta,\theta'')$, corresponds to part of what is meant by a good interval estimator; in particular, small values of $a(u,\theta,\theta'')$ for $u > \theta$, which indicate desirable properties of accuracy or shortness for an upper estimator θ'', indicate corresponding shortness properties for an interval estimator $J = (\theta',\theta'')$.

The merits of any interval estimator J depend upon the following considerations in suitable combination.

(a) The probability should be suitably high that the indicated conclusions ("θ lies in $[\theta', \theta'']$") are correct, for each possible true value of θ. That is, the confidence coefficient should have a suitably high value; or, more generally, for each θ, the sum of the location functions $a(\theta-,\theta,J)$ and $a(\theta+,\theta,J)$ should be suitably low. As with point estimators, it seems desirable to avoid, as far as possible and convenient in the development of a general theory, any step which corresponds to a tacit judgment that errors of overestimation and underestimation are necessarily comparable either

qualitatively or quantitively. Hence the present specification will be given the form: Each of the location functions $a(\Theta-,\Theta,J)$, $a(\Theta+,\Theta,J)$ should have suitably small values, for each Θ.

(b) Given the location functions of an interval estimator (and, hence, given the probability $1-a(\Theta-,\Theta,J) - a(\Theta+,\Theta,J)$ of correct conclusions, for each Θ), the indicated conclusions should when correct be as strong and informative as possible. That is, for each Θ, the conditional distributions of $\Theta'(X)$ and $\Theta''(X)$, given that $\Theta(X) \leqq \Theta \leqq \Theta''(X)$, should be concentrated as close to Θ as possible. (In terms of the conditional bivariate distribution of $(\Theta'(X), \Theta''(X)$, this means concentration close to the point (Θ,Θ).) These desirable shortness properties of J correspond to suitably small values, for each Θ, of $a(u,\Theta,\Theta'')$ for each $u > \Theta$ and of $a(u,\Theta,\Theta')$ for each $u < \Theta$.

(c) Given that one of the conclusions indicated by J is incorrect, it should be misleading in the smallest possible degree. (The remarks on property (c) of the preceding section are also applicable here.) These desirable closeness properties of J correspond to suitably small values of $a(u,\Theta,J)$ for each Θ and each $u \neq \Theta$; that is, suitably small values of $a(u,\Theta,\Theta')$ for $u > \Theta$ and of $a(u,\Theta,\Theta'')$ for $u < \Theta$.

To represent all of the properties considered for interval estimators, we define the risk curves of each interval estimator $J = (\Theta',\Theta'')$, at each Θ, as the pair of functions $[a(u,\Theta,\Theta'), a(u,\Theta,\Theta'')]$ of $u(u \neq \Theta)$, i.e. the risk curves of Θ' and of Θ''. Thus the risk curves of J at Θ are a representation of the bivariate cumulative distribution function of $\Theta'(X)$ and $\Theta''(X)$ when Θ is true.

These considerations lead us to formulate the following basic definitions: An interval estimator $J = (\theta', \theta'')$ will be called at least as good as another $J^* = (\theta^*, \theta^{**})$ if θ' is at least as good as θ^* and θ'' is at least as good as θ^{**} in the sense defined for point estimators in Section 2 above. Similarly, J will be called better than J^* if it is at least as good as J^* and also θ' is better than θ^* and/or θ'' is better than θ^{**}. J will be called admissible if no other interval estimator is better. Complete classes are defined in the usual way.

If two interval estimators have different location functions, they are not comparable (neither is at least as good as the other); this follows immediately from the corresponding property for point estimators. A simple sufficient condition for admissibility of $J = (\theta', \theta'')$ is that θ' and θ'' be admissible point estimators.

5. Confidence curve estimators. The selection of an estimator of one of the above kinds for purposes of informative inference, including typical applications in scientific research, is generally admitted to involve elements of choice which are in some degree arbitrary. Such elements include the choice of a particular confidence level for an interval estimator, and the choice of location functions for an interval estimator with given confidence coefficient. In addition, a point estimate is sometimes desired along with an interval. Such considerations and related ones have led to proposals for use simultaneously of a point estimator and a set of confidence limit or interval estimators having various confidence coefficients. Such estimators may be regarded as a modern formulation of a long-standing practice of reporting

estimates in the form $\theta^* \pm k\sigma_{\theta^*}$, where k is some constant and $\sigma^2_{\theta^*} = \text{Var}\ (\theta^*(X))$. The latter form may be interpreted as an ordered set of three point estimators. For example, if $\theta^*(X)$ has a normal distribution with a known constant variance, and $k = 1$, then the "estimator" $\theta^*(x) \pm k\sigma_{\theta^*}$ may be written as the ordered set of estimators

$$[\theta^*(x) - \sigma_{\theta^*},\ \theta^*(x), \theta^*(x) + \sigma_{\theta^*}] \equiv [\theta(x,.84), \theta(x,.5),\ \theta(x,.16)].$$

Estimates of this "omnibus" kind can be interpreted flexibly but validly, in any context of application for informative inferences, in the ways customary for (a) point estimates such as $\theta(x,.5)$, (b) confidence limits such as $\theta(x,.84)$ and $\theta(x,.16)$, and (c) confidence intervals such as $[\theta(x,.84),\ \theta(x,.16)]$.

Tukey [10] proposed that for typical general purposes it would be advantageous to use a set of five point estimators at standard levels: $\theta(x,\alpha)$, with $\alpha = 2\frac{1}{2}\%$, $16\frac{2}{3}\%$, 50%, $83\frac{1}{3}\%$, and $97\frac{1}{2}\%$. Cox [11] proposed use of the full continuous family of confidence limits $\theta(x,\alpha)$, $0 \leqq \alpha \leqq 1$. Such an omnibus estimator includes formally, as elements, not only confidence limits at all levels and a median-unbiased point estimator, but also median-unbiased confidence intervals at all levels. Whether such estimators should be used in practice, rather than more standard methods, is a matter of judgment and taste which can perhaps be decided best in specific contexts of application. It is often convenient, as will be illustrated below, to discuss estimation theory and techniques for estimators of this omnibus form, since such discussion includes conveniently and compactly a treatment of estimators of the various

kinds mentioned.

Any such estimator, consisting of a specified set of confidence limit estimators $\Theta(x,\alpha)$, α in some specified subset of the closed unit interval (possibly the whole interval), ordered in the sense that $\alpha < \alpha'$ implies $\Theta(x,\alpha) \geq \Theta(x,\alpha')$ for each x in S, will be called a confidence curve estimator. We shall usually consider the inclusive case, $0 \leqq \alpha \leqq 1$, so as to include formally all other cases. In many problems it is convenient to give such estimators a form which can be reported graphically: if for each $x \in S$, $\Theta(x,\alpha)$ increases continuously from $\underline{\theta}$ to $\overline{\theta}$ as α decreases from 1 to 0, then we define the confidence curve estimator $c(\theta,x)$, for each $x \in S$, as the continuous curve (function of $\theta \in \overline{\Omega}$)

$$c(\theta,x) = \min [\alpha, 1-\alpha | \Theta(x,\alpha) = \theta] .$$

For example, if X is normally distributed with unit variance and mean θ, then the confidence curve estimator of θ is

$$c(\theta,x) = \begin{cases} \Phi(\theta - x), & -\infty \leqq \theta \leqq x, \\ 1 - \Phi(\theta - x), & x \leqq \theta \leqq \infty; \end{cases}$$

for any observed value x, the estimate $c(\theta,x)$ can be described by a more or less complete sketch of its graph when convenient. Such estimates are illustrated in a number of examples in Section 9 below.

The definitions of admissibility and of complete classes for confidence curve estimators parallel those above for confidence

interval estimators. A simple sufficient (but not, in general, necessary) condition that a confidence curve estimator be admissible is that for each α, its element $\theta^*(x,\alpha)$ be an admissible point estimator. In problems for which there exists a uniformly best confidence limit estimator for each confidence coefficient, this condition is necessary as well as sufficient, and there is a unique (a.e.) admissible confidence curve estimator which consists simply of the family of these best confidence limit estimators.

6. Elementary theory of admissible point estimators. An important part of the general theory of admissible point estimators, and of corresponding practical techniques of estimation, can be developed conveniently by an essentially elementary use of the theory of tests of one-sided hypotheses as originated by Neyman and Pearson and as extended (by simple use of their Fundamental Lemma) to generate a variety of admissible tests of such hypotheses. In problems for which uniformly best one-sided tests exist, the complete theory of admissible estimators is obtained in this way; for other problems, the development of the remaining parts of the theory requires more general methods introduced in Section 10 below.

For each θ_0 in Ω, we consider two one-sided testing problems: (a) the problem of testing the hypothesis $H(\theta_0)$: $\theta \leq \theta_0$ (against the general alternative $H'(\theta_0)$: $\theta > \theta_0$); and (b) the problem of testing $H(\theta_0-)$: $\theta < \theta_0$ (against the general alternative $H'(\theta_0-)$: $\theta \geq \theta_0$). In case θ_0 is a minimum value in Ω, consideration of $H(\theta_0-)$ is to be ommitted; if θ_0 is a maximum in Ω, $H(\theta_0)$ is omitted.

Any given point estimator $\theta^* = \theta^*(x)$ of θ can be used in the following way to define a test of each of the hypotheses mentioned:

Accept the hypothesis if and only if the observed value $\theta^*(x)$ is consistent with the hypothesis. Such a test of the hypothesis $H(\theta_0)$ has the acceptance region $A(\theta_0) = \{x|\theta^*(x) \leqq \theta_0\}$; such a test of $H(\theta_0-)$ has acceptance region $A(\theta_0-) = \{x|\theta^*(x) < \theta_0\}$. If $\theta_1 < \theta_2$, then $A(\theta_1-) \subset A(\theta_1) \subset A(\theta_2-) \subset A(\theta_2)$; for brevity, we shall say that such a sequence of sets $A(\theta)$ is nondecreasing in θ, with the understanding the argument θ may take a value $(\theta-)$ which is considered smaller than θ and larger than $\theta-\varepsilon$ for each positive ε.

Such a test of $H(\theta_0-)$ has probabilities of errors of Type I given by

$$1 - \text{Prob}\,(A(\theta_0-)|\theta) = a(\theta_0,\theta,\theta^*) \text{ for each } \theta < \theta_0,$$

and of Type II given by

$$\text{Prob}\,(A(\theta_0-)|\theta) = a(\theta_0-,\theta,\theta^*) \text{ for each } \theta \geqq \theta_0 .$$

Such a test of $H(\theta_0)$ has probabilities of errors of Type I given by

$$1 - \text{Prob}\,(A(\theta_0)|\theta) = a(\theta_0 + ,\theta,\theta^*) \text{ for each } \theta \leqq \theta_0 ,$$

and of Type II given by

$$\text{Prob}\,(A(\theta_0)|\theta) = a(\theta_0,\theta,\theta^*) \text{ for each } \theta > \theta_0 .$$

Thus each of the error-probabilities $a(u,\theta,\theta^*)$, upon which depend the admissibility of any given point estimator θ^*, appears as an error-probability of a test of a one-sided hypothesis based upon use of θ^*. These relationships provide the following simple sufficient condition for admissibility of a point estimator.

Lemma 1. For any specified family of probability density functions $f(x,\theta)$ (with respect to an underlying σ-finite measure $\mu(x)$ defined

on the sample space $S = \{x\}$), $\theta \in \Omega$ (a subset of the real line), a given estimator $\theta^* = \theta^*(x)$ (any measurable function taking values in the closure $\bar{\Omega}$ of Ω) is admissible if each of the acceptance regions $A(\theta_0)$, $A(\theta_0-)$, based on θ^* as defined above, gives an admissible test of the corresponding one-sided hypotheses $H(\theta_0)$, $H(\theta_0-)$ defined above.

Proof: (A test is called admissible if no other test has all error-probabilities at least as small, with at least one strictly smaller.) If θ^* satisfies the assumptions of the Lemma but is inadmissible, let θ^{**} be an estimator better than θ^*. Then $a(\theta_0,\theta,\theta^{**}) \leqq a(\theta_0,\theta,\theta^*)$ for each $\theta \in \Omega$ and each $\theta_0 \neq \theta$, and the inequality is strict for some $\theta = \theta' \in \Omega$ and some $\theta_0 = \theta_0' \in \bar{\Omega}$, $\theta_0' \neq \theta'$. Assume for definiteness that $\theta_0' > \theta'$ (the other case can be discussed in the same way). Then the acceptance region $\{x \mid \theta^{**}(x) < \theta_0'\}$ gives a better test of the hypothesis $H(\theta_0'-)$ than does $\{x \mid \theta^*(x) < \theta_0'\}$. This contradicts the assumed admissibility of the test based on the latter region, completing the proof.

Many estimators of interest can be conveniently investigated theoretically and constructed practically by the device of using as indicated below a function $v(x,\theta)$, defined for each sample point x and each $\theta \in \Omega$. If, for each fixed θ, $v(x,\theta)$ is a measurable function of x, it is a <u>statistic</u>; and as θ varies, $v(x,\theta)$ represents a family of statistics. We term such a function v a <u>quasistatistic</u>.

<u>Corollary 1</u>. A sufficient condition for admissibility of an estimator $\theta^*(x)$ is that it be defined, for each x, as the solution θ of the equation $v(x,\theta) = 0$, where v is a quasistatistic such that:

(a) For each x in S, $v(x,\theta) = 0$ holds for a unique θ in $\bar{\Omega}$.

(b) If $\theta_1 < \theta_2$ and θ_1, θ_2 are in $\bar{\Omega}$, then $\{x|v(x,\theta_1) \leqq 0\} \subset \{x|v(x,\theta_2) < 0\}$.

(A simple sufficient condition for (b) is that for each x, $v(x,\theta)$ be nonincreasing in θ.)

(c) For each θ_0 in Ω, the acceptance regions $\{x|v(x,\theta_0) \leqq 0\}$ and $\{x|v(x,\theta_0) < 0\}$ are admissible respectively for testing the one-sided hypotheses $H(\theta_0)$ and $H(\theta_0-)$.

Proof: If $v(x,\theta)$ satisfies the stated conditions, the conclusion follows immediately from Lemma 1 upon observing that

$$\{x|v(x,\theta_0) \leqq 0\} = \{x|\theta^*(x) \leqq \theta_0\} \text{ and } \{x|v(x,\theta_0) < 0\} = \{x|\theta^*(x) < \theta_0\}$$

When an estimator θ^* is defined implicitly, by use of a quasi-statistic $v(x,\theta)$, as the solution θ of the equation $v(x,\theta) = 0$, in applications it is not necessary to have an explicit formula for $\theta^*(x)$ since for any observed sample point x it suffices merely to determine the corresponding root θ of the defining equation; and in the cases of many such estimators of practical and theoretical interest, no explicit formula for $\theta^*(x)$ is available. The preceding lemma shows that basic qualitative properties of efficiency can be established for such estimators without use of any explicit formula for $\theta^*(x)$. Their quantitive properties can also be determined without such explicit formulas: Since $v(x,u) < 0$ is equivalent to $\theta^*(x) < u$, and $v(x,u) = 0$ is equivalent to $\theta^*(x) = u$, we have

$$a(u,\theta,\theta^*) = \begin{cases} \text{Prob } [\theta^*(X) \leqq u|\theta] = \text{Prob } [v(X,u) \leqq 0|\theta] \text{ for } u < \theta \\ \text{Prob } [\theta^*(X) \geqq u|\theta] = \text{Prob } [v(X,u) \geqq 0|\theta] \text{ for } u > \theta. \end{cases}$$

Thus all quantitative properties of such estimators θ^* can be determined, when convenient, by determining
Prob $[v(X,u) \leqq 0|\theta]$ and Prob $[v(X,u) = 0|\theta]$ for each $u \neq \theta$.

Some theoretical properties of such estimators are also conveniently treated in terms of the c.d.f's. of v. For example, if for each n = 1, 2,..., θ_n^* is an estimator determined by a quasi-statistic $v_n = v_n(x_n,\theta)$, then the condition that the sequence of estimators θ_n^* be <u>consistent</u> (that is, that $\text{Lim}_n\ a(u,\theta,\theta_n^*) = 0$, for each $\theta \in \Omega$ and each $u \neq \theta$), can be stated, and in many cases conveniently proved, in the form: Lim_n Prob $[v_n(X_n,u) \leqq 0|\theta] = 0$ or 1, according as $u < \theta$ or $u > \theta$, for each $\theta \in \Omega$.

For estimation by confidence intervals or confidence curves, it is sometimes convenient to employ a family of quasistatistics. Suppose that for each of several values of an index α, $v(x,\theta,\alpha)$ is a quasistatistic which determines as above an estimator $\theta(x,\alpha)$, and that, for each x in S, $\theta(x,\alpha)$ is decreasing in α. Then for any pair of values of α, $\alpha' > \alpha''$, the pair of estimators $[\theta(x,\alpha'), \theta(x,\alpha'')] = J(x)$ is an interval estimator of θ, whose quantitative properties may be investigated in terms of the distributions of $v(X,u,\alpha)$ as indicated above, and whose admissibility can in some cases be established by direct application of Corollary 1 to $v(x,\theta,\alpha')$ and $v(x,\theta,\alpha'')$. A case of interest is that in which $\alpha = \text{Prob } [v(X,\theta,\alpha) \leqq 0|\theta] \equiv \text{Prob } [v(X,\theta,\alpha) < 0|\theta]$ for each

α, $0 \leqq \alpha \leqq 1$, and each $\theta \in \Omega$. Then the family of estimators $\theta(x,\alpha)$ constitutes a confidence curve estimator of θ (assuming again that $v(x,\theta,\alpha)$ is decreasing in α); this estimator is admissible if for each α the quasistatistic $v(x,\theta,\alpha)$ satisfies the assumptions of Corollary 1. Examples of such estimators, and of convenient techniques for their computation and presentation, are given below.

7. Uniformly best estimators. Let $\theta^*(x)$ be any estimator of $\theta \in \Omega$. θ^* will be called a uniformly best estimator of θ if, among all estimators with the same location functions $a(\theta-,\theta)$, $a(\theta+,\theta)$, θ^* has uniformly minimum error-probabilities $a(u,\theta)$. Since the $a(u,\theta)$'s are error-probabilities of tests of one-sided hypotheses $H(\theta_0-)$, $H(\theta_0)$, $\theta_0 \in \Omega$, with respective acceptance regions $A(\theta_0-) = \{x|\theta^*(x) < \theta_0\}$, $A(\theta_0) = \{x|\theta^*(x) \leqq \theta_0\}$, a necessary condition for θ^* to be a uniformly best estimator is that $f(x,\theta)$ and Ω admit uniformly best tests of the hypotheses $H(\theta_0-)$, $H(\theta_0)$, of respective sizes $a(\theta_0-, \theta_0, \theta^*)$, $1 - a(\theta_0+, \theta_0, \theta^*)$, $\theta_0 \in \Omega$.

It is well known [12] that uniformly best one-sided tests of all sizes exist if and only if there exists a sufficient statistic $t(x)$ with the monotone likelihood ratio (m.l.r.) property, in which case each best test may be obtained by use of an acceptance region of the form

$$A(\theta_0-) = \{(x,y)|\ z(t(x),y,\theta_0) \leqq a(\theta_0-,\theta_0)\} \text{ or}$$

$$A(\theta_0) = \{(x,y)|\ z(t(x),y,\theta_0) \leqq 1-a(\theta_0+,\theta_0)\},$$

where Y is the observed value of a uniformly distributed auxiliary

randomization variable y, $0 \leq Y < 1$, and Z is the continuous probability integral transform of Y:
$z(t(x),y,\theta) = yF(t(x),\theta) + (1-y)F(t(x)-,\theta)$, where
$F(t,\theta) = \text{Prob}\{t(X) \leq t|\theta\}$. If such a sufficient statistic t(x) exists, then a simple sufficient condition for admissibility of an estimator θ^* is clearly that θ^* be a non-decreasing function of t(x); for then $A(\theta_o-) = \{x|\theta^*(t(x)) < \theta_o\}$ and $A(\theta_o) = \{x|\theta^*(t(x)) \leq \theta_o\}$ are uniformly best one-sided tests. If such a statistic t(x) has a discrete distribution on a subset of the integers, then t(x) + y is another sufficient statistic having the monotone likelihood ratio property, and having a continuous c.d.f. under each θ; as above, a simple sufficient condition for admissibility of an estimator θ^* is that it be a non-decreasing function of t(x) + y.

More generally, let θ^* be <u>any</u> estimator, let $G(\theta) = \text{Prob}\{\theta^*(X) \leq \theta|\theta\}$, let $G(\theta-) = \text{Prob}\{\theta^*(X) < \theta|\theta\}$, let $F(t,\theta) = \text{Prob}\{t(X) \leq t|\theta\}$, where t(x) is a sufficient statistic with the m.l.r. property, and as above let $z(t(x),y,\theta) = yF(t(x),\theta) + (1-y)F(t(x)-,\theta)$. Consider the quasi-statistic $v = v(x,y,\theta) = z(t(x),y,\theta) - G(\theta)$. For each θ_o, $A(\theta_o) = \{(x,y)|v(x,y,\theta_o) < 0\}$ is clearly a uniformly best acceptance region for testing $H(\theta_o)$ at level $1-G(\theta_o) = a(\theta_o+,\theta_o,\theta^*)$. Consider the quasistatistic $v' = v'(x,y,\theta) = z(t(x),y,\theta)-G(\theta-) \leq v + [G(\theta) - G(\theta-)]$. For each θ_o, $A(\theta_o-) = \{(x,y)|v'(x,y,\theta_o) < 0\}$ is clearly a uniformly best acceptance region for testing $H(\theta_o-)$; at $\theta = \theta_o$ it has Type II error probability $G(\theta_o-) = a(\theta_o-,\theta_o,\theta^*)$.

To verify that these acceptance regions constitute a sequence of sets which is nondecreasing in θ in the sense defined in Section 6, we note that obviously $A(\theta_0-) \subset A(\theta_0)$, and we proceed to prove that $\theta_1 < \theta_2$ implies $A(\theta_1) \subset A(\theta_2-)$: Assume that $(x',y') \in A(\theta_1)$; but $(x',y') \notin A(\theta_2-)$; then $z' \equiv z(t(x'),y',\theta_1) < G(\theta_1)$ and $z'' \equiv z(t(x'),y',\theta_2) \geqq G(\theta_2-)$. A best test of $H(\theta_1)$ of size $(1-z')$ (the test which rejects when $z(t(x),y,\theta_1) \geqq z'$) has maximum power at $\theta = \theta_2$, namely $1-z''$; the test with acceptance region $\{x|\theta^*(x) \leqq \theta_1\}$ has size $1 - G(\theta_1) < (1-z')$ and hence has power $\text{Prob}\{\theta^*(X) \leqq \theta_1|\theta_2\} < 1 - z''$ Hence $z'' < \text{Prob}\{\theta^*(X) \leqq \theta_1|\theta_2\} \leqq \text{Prob}\{\theta^*(X) < \theta_2\} = G(\theta_2-)$, a contradiction which proves that $A(\theta_1) \subset A(\theta_2-)$.

For each (x,y), let $\theta^{**} = \theta^{**}(x,y)$ be defined by $\theta^{**}(x,y) = \inf\{\theta|\theta \in \Omega, (x,y) \in A(\theta)\}$. Then θ^{**} is a nondecreasing function of $t(x)$ and of y, and is a uniformly best estimator having the same location functions as the arbitrarily given θ^*. If each best test is admissible, then θ^{**} is admissible, and hence is strictly better than θ^* or else it is equivalent to θ^*. These considerations establish the following

Lemma 2. If the family of density functions $f(x,\theta)$, $\theta \in \Omega$, admits a sufficient statistic $t = t(x)$ having the monotone likelihood ratio property, then an essentially complete class of estimators is constituted by estimators of the form $\theta^* = \theta^*(t,y)$, any nondecreasing function of t and of y, where y is an observed value of an auxiliary randomization variable Y having under each θ the same uniform distribution on the unit interval $0 \leqq y < 1$, and such that $t' < t''$ implies $\theta^*(t',y') \leqq \theta^*(t'',y'')$ for all y',y''.

If $t(x)$ has a continuous c.d.f., for each θ, then estimators of this form but not depending upon y constitute an essentially complete class of estimators.

8. Score quasistatistics and generalized maximum likelihood estimators.

For a given family $f(x,\theta)$, $\theta \in \Omega$, let $\theta_1(\theta)$, $\theta_2(\theta)$ be two functions defined on Ω, taking values in $\bar{\Omega}$, and satisfying $\theta_1(\theta) < \theta_2(\theta)$ and $\theta_1(\theta) \leqq \theta \leqq \theta_2(\theta)$ for $\theta \in \Omega$. Then for each $\theta' \in \Omega$, a best test of $H_1: \theta = \theta_1(\theta')$ against $H_2: \theta = \theta_2(\theta')$ is one which accepts H_1 when the quasistatistic

$$S(x,\theta_1(\theta),\theta_2(\theta)) = [\log f(x,\theta_2(\theta)) - \log f(x,\theta_1(\theta))]/[\theta_2(\theta)-\theta_1(\theta)]$$

satisfies $S(x,\theta_1(\theta'), \theta_2(\theta')) \leq G(\theta',\alpha(\theta'))$, where $G(\theta,\alpha(\theta'))$ is a constant such that $\alpha(\theta')$ is the probability, when θ' is true, that this inequality will be satisfied. For many problems the functions $\theta_1(\theta)$, $\theta_2(\theta)$, and $\alpha(\theta)$ can be chosen so that the generalized score quasistatistic $v(x,\theta) = S(x,\theta_1(\theta),\theta_2(\theta)) - G(\theta,\alpha(\theta))$, $\theta \in \Omega$, satisfies the conditions of Corollary 1 and hence defines an admissible estimator $\theta^*(x)$ as the solution θ of the equation $v(x,\theta) = 0$. If, for example, Prob $\{v(X,\theta) = 0|\theta\} \equiv 0$ for $\theta \in \Omega$, and the set $\{x|f(x,\theta) > 0\}$ is independent of $\theta \in \Omega$, then each acceptance region $\{x|v(x,\theta) \leqq 0\}$ gives a best test which is essentially unique (a.e. P_θ, $\theta \in \Omega$), and hence admissible for testing $H(\theta)$ and $H(\theta-)$.

Again, as $\theta_2(\theta) - \theta_1(\theta) \longrightarrow 0$, $S(x,\theta_1(\theta), \theta_2(\theta)) \longrightarrow S(x,\theta)$

$$= \frac{\partial}{\partial\theta} \log f(x,\theta),$$

if the derivative exists at each x, for each $\theta \in \Omega$; consider as

above the <u>(locally-best) score quasistatistic</u>
$v(x,\theta) = S(x,\theta) - G(\theta,a(\theta))$. Again, if this $v(x,\theta)$ satisfies the conditions of Corollary 1, then an admissible estimator $\theta^{*}(x)$ is defined as the solution θ of the equation $v(x,\theta) = 0$. It is well known that, under a mild regularity condition, an acceptance region $\{x|v(x,\theta) \leqq 0\}$ gives a locally-best test of $H(\theta)$ and of $H(\theta-)$; under additional mild restrictions, such as those mentioned above, these tests are also admissible. The case $G(\theta,a(\theta)) \equiv 0$, $\theta \in \Omega$, determines (through the equation $S(x,\theta) = 0$) the maximum likelihood estimator $\hat{\theta}(x)$, which is thus shown to be admissible (and to be locally-best, i.e. to minimize $a(u,\theta)$ for θ near u, among all estimators with the same location functions) provided that $v(x,\theta) = S(x,\theta)$ satisfies the conditions mentioned. Estimators of this form were proposed by Tukey [10] on different theoretical grounds in connection with the methods discussed in Section 5 above.

Estimators defined by use of the various score quasistatistics mentioned may be called <u>generalized maximum likelihood estimators</u>.

If $S(x,\theta)$ has (or may have) discontinuous distributions, it can be replaced, as may be desired at least for some theoretical purposes, by its continuous probability integral transform

$$a(x,y,\theta) = y \cdot \text{Prob}\,[S(X,\theta) \leqq S(x,\theta)|\theta],$$
$$+ (1-y) \cdot \text{Prob}\,[S(X,\theta) < S(x,\theta)|\theta],$$

where y is the observed value of Y, an auxiliary randomization variable having, for each θ, the same uniform density on $0 \leqq y < 1$.

Then for each θ, $\alpha(\theta)$ may be prescribed arbitrarily, and the statistic

$$v(x,y,\theta,\alpha(\theta)) = a(x,y,\theta) - \alpha(\theta)$$

has a continuous distribution and takes negative values with probability $\alpha(\theta)$. In suitable problems, with suitable choices of $\alpha($ the quasistatistic v so defined will satisfy the conditions of Corollary 1. The same treatment can be applied to the form $S(x,\theta_1(\theta),\theta_2(\theta))$. To avoid technicalities of little intrinsic interest, we discuss the case in which such randomization is not used.

If Prob $\{v(x,\theta) = 0|\theta\} = 0$ for each $\theta \in \Omega$, then each such estimator has the location functions $a(\theta-,\theta) \equiv 1 - a(\theta+,\theta) \equiv \alpha(\theta)$. If $\alpha(\theta) \equiv a$, a constant, such an estimator is a confidence limit; if $\alpha(\theta) \equiv 1/2$, such an estimator is a median-unbiased point estimator. In the important case that $X = (Y_1,\ldots Y_n)$, a sample of independent observations Y_i, we have $S(X,\theta) = \sum_{i=1}^{n} S(Y_i,\theta)$; the normal approximation (based on the Central Limit Theorem)

$$a(\theta-,\theta,\hat{\theta}) = \text{Prob}\{S(X,\theta) < 0|\theta\} \doteq \Phi(0) = 1/2$$

(using that $E(S(X,\theta)|\theta) \equiv 0$) is often close; hence in such cases the maximum likelihood estimator $\hat{\theta}(x)$ is approximately median-unbiased. If $S(X,\theta)$ has a symmetrical distribution under θ, then clearly $\hat{\theta}$ is exactly median-unbiased.

In some cases, as illustrated below, a family of score quasi-statistics, e.g.

$$v(x,\theta,\alpha) = S(x,\theta) - G(\theta,\alpha),\ 0 \leqq \alpha \leqq 1,$$

or

$$v(x,\theta,\alpha) = S(x,\theta_1(\theta),\theta_2(\theta)) - G(\theta,\alpha),\ 0 \leqq \alpha \leqq 1,$$

can be used to determine admissible confidence curve estimators $\theta(x,\alpha)$, $0 \leqq \alpha \leqq 1$, as solutions of equations $v(x,\theta,\alpha) = 0$.

Estimators based on score quasistatistics have direct usefulness, which is enhanced by the simplicity of their theory and of the practical techniques for their use. In addition they are of special theoretical interest, due to their relations to the asymptotic theory and techniques of maximum likelihood estimation; they generalize and justify these techniques in an exact sense. The following considerations lend them further intrinsic interest: For any given problem of estimation of θ, consider the class of estimators having specified location functions $a(\theta-,\theta)$, $a(\theta+,\theta)$. For each $\theta \in \Omega$ and each $u \neq \theta$, $u \in \Omega$, let $\underline{a}(u,\theta) = \min_{\theta^*} a(u,\theta,\theta^*)$, where for $u > \theta$ the minimum is taken over all estimators such that $a(\theta+,\theta,\theta^*) = a(\theta+,\theta)$, and for $u < \theta$ the minimum is taken over all estimators such that $a(\theta-,\theta,\theta^*) = a(\theta-,\theta)$. Then $\underline{a}(u,\theta)$ is the <u>envelope risk curve</u> (i.e. the minimum of the respective ordinates of risk curves) for the class of estimators with the given location functions. For each (u,θ), it is possible to attain $\underline{a}(u,\theta)$ in the following sense: if $u > \theta$, the relatively trivial estimator which takes the value θ with probability $1 - a(\theta+,\theta)$ when θ is true, and which takes the value u otherwise, and which minimizes $a(u,\theta,\theta^*)$ subject to these conditions, is equivalent to a best test between the simple hypotheses θ and u, of the indicated size; such a test

can be based on the score statistic $S(x,u,\theta)$; similar remarks apply to the case $u < \theta$. Each such single statistic $S(x,u,\theta)$ can be embedded, as an element, in a score quasistatistic $S(x,\theta_1(\theta),\theta_2(\theta))$ for $\theta \in \Omega$; it may or may not be possible to define by use of this quasistatistic an estimator which has the specified location functions. An estimator can attain $\underline{a}(u,\theta)$ uniformly in (u,θ) only in problems having the special structure described in Section 7 above, for which uniformly best estimators exist. In other problems, some estimators defined by generalized score quasistatistic attain $\underline{a}(u,\theta)$ at some but not all (u,θ). In all problems, the computation of $\underline{a}(u,\theta)$ requires calculations of probabilities of events defined by score statistics $S(x,u,\theta)$; and the possibility of its attainment by some estimator at specified points (u,θ) is related to the existence of suitable score quasistatistics.

8.1 Large-sample approximations.

If $x = (y_1,\ldots y_n)$ is a sample of n independent identically distributed observations (non-identical distributions can be discussed similarly), $S(x,\theta_1(\theta),\theta_2(\theta)) = \sum_{i=1}^{n} S(y_i,\theta_1(\theta),\theta_2(\theta))$. Let $\mu(u,\theta) = E[S(Y_1,\theta_1(u))|\theta]$ and $\sigma^2(u,\theta) = \mathrm{Var}\,[S(Y_1,\theta_1(u),\theta_2(u))|\theta]$ exist for each $\theta,u \in \bar{\Omega}$. We allow $\theta_1(\theta) = \theta_2(\theta) = \theta$ here, taking $S(X,\theta,\theta) \equiv S(X,\theta)$ in this case, and assume that $\theta_1(\theta),\theta_2(\theta)$ are fixed, while n may vary, in the present discussion.

In the special case $v_n(x,\theta) = \sum_{i=1}^{n} S(y_i,\theta)$, which determines the maximum likelihood estimator $\hat{\theta}_n(x)$ as the solution θ of $v_n(x,\theta) = 0$, we have by Khintchine's Theorem (even if $\sigma^2(u,\theta)$'s do not exist) that $\frac{1}{n} v_n(X,u)$ converges in probability to $\mu(u,\theta)$ when

θ is true. If $u' < \theta < u''$ implies $\mu(u',\theta) < \mu(\theta,\theta) \equiv 0 < \mu(u'',\theta)$, then $\operatorname{Lim}_n a(u,\theta,\hat{\theta}_n) = 0$ for $u \neq \theta$; that is, $\hat{\theta}_n$ is consistent.

Returning to the general case, for large n the Central Limit Theorem gives the normal approximation to the distributions of

$$v_n(X,u,\alpha) = \sum_{i=1}^{n} S(Y_i,\theta_1(u),\theta_2(u)) - G_n(u,\alpha):$$

$$\text{Prob}\left\{v_n(X,u,\alpha) \leqq 0|\theta\right\} \doteq \Phi\left(\frac{G_n(u,\alpha)-n\mu(u,\theta)}{\sqrt{n}\,\sigma(u,\theta)}\right);$$

and for $u = \theta$, the approximate determination of $G_n(\theta,\alpha)$:

$$\alpha \doteq \Phi\left(\frac{G_n(\theta,\alpha)}{\sqrt{n}\sigma(\theta,\theta)}\right), \text{ or } G_n(\theta,\alpha) \doteq \sqrt{n}\sigma(\theta,\theta)\Phi^{-1}(\alpha),$$

which in the preceding formula gives

$$\text{Prob}\left\{v_n(X,u) \leqq 0|\theta\right\} \doteq \Phi\left(-\sqrt{n}\,\frac{\mu(u,\theta)}{\sigma(u,\theta)} + \frac{\sigma(\theta,\theta)}{\sigma(u,\theta)}\,\Phi^{-1}(\alpha)\right).$$

For the maximum likelihood estimator, $G_n = 0$, corresponding to $\alpha = \frac{1}{2}$ in these formulae. Thus the risk curves of the confidence limit estimator $\theta^* = \theta_n(x,\alpha)$ determined by $v_n(x,\theta,\alpha) = 0$ are approximately

$$a(u,\theta,\theta_n(\cdot,\alpha)) = \begin{cases} \Phi(h(u,\theta,\alpha,n)), & u < \theta, \\ 1 - \Phi(h(u,\theta,\alpha,n)), & u > \theta, \quad 0 < \alpha < 1, \end{cases}$$

where

$$h(u,\theta,\alpha,n) = -\sqrt{n}\,\frac{\mu(u,\theta)}{\sigma(u,\theta)} + \frac{\sigma(\theta,\theta)}{\sigma(u,\theta)}\,\Phi^{-1}(\alpha).$$

Here the sufficient (and necessary) condition for consistency of $\theta_n(x,\alpha)$, for a fixed α, $0 < \alpha < 1$, is again that $u' < \theta < u''$ imply $\mu(u',\theta) < 0 < \mu(u'',\theta)$.

The verification of the conditions of Corollary 1, for a given $v(x,\theta)$, is sometimes difficult. Large-sample approximations are of some theoretical and practical help in this connection. For example for a locally best confidence limit estimator $\theta(x,\alpha)$, where $x = (y_1,\ldots y_n)$ and the Y_i's are independent and identically distributed, we have as above

$$G_n(\theta,\alpha) \triangleq \sqrt{n}\sigma(\theta,\theta)\,\Phi^{-1}(\alpha)\ ,$$

and we take

$$v_n(x,\theta,\alpha) = S(x,\theta) - \sqrt{n}\sigma(\theta,\theta)\Phi^{-1}(\alpha)\ .$$

If $S(x,\theta)$ satisfies the conditions of Corollary 1 (i.e. if for each x the maximum likelihood estimator $\hat{\theta}(x)$ is determined as the root θ of $S(x,\theta) = 0$), and is decreasing in θ for each x, then:

(A) If $\sigma(\theta,\theta)$ is constant (this is the case in some examples in the following Section, but not in most examples), then $v_n(x,\theta,\alpha)$ is also decreasing, as required by Corollary 1.

(B) If $\sigma(\theta,\theta)$ is decreasing or increasing at $\theta = \theta'$, then for a fixed x and α sufficiently near 0 or 1, either

$$-n\sigma(\theta,\theta)\Phi^{-1}(\alpha) \text{ or } -n\sigma(\theta,\theta)\Phi^{-1}(1-\alpha) \equiv n\sigma(\theta,\theta)\Phi^{-1}(\alpha)$$

will be increasing more rapidly than $S(x,\theta)$ is decreasing at $\theta = \theta'$, so that $v_n(x,\theta,\alpha)$ and $v_n(x,\theta,1-\alpha)$ cannot both be decreasing in θ at θ'.

(C) On the other hand, for any fixed α, $0 < \alpha < 1$, since

$$v_n(x,\theta,\alpha) = \sum_{i=1}^{n} [S(y_i,\theta) - \frac{\sigma(\theta,\theta)}{\sqrt{n}} \Phi^{-1}(\alpha)],$$

a sufficient condition for v_n to be decreasing in θ is that

$$S(y_1,\theta) - \frac{\sigma(\theta,\theta)}{\sqrt{n}} \Phi^{-1}(\alpha)$$

be decreasing in θ, for all values of y_1. Clearly as n increases, this condition becomes a less restrictive one, being in general satisfied for a wider range of values of α.

8.2 Local approximations for locally best estimators.

In cases where there exist precise estimators, that is estimators whose risk curves are small except for u very near θ, it is natural to center attention on small neighborhoods of the possible true values θ, and to consider estimators whose risk curves are relatively small in such neighborhoods, such as those based on score quasistatistics with $\theta_2(\theta)-\theta_1(\theta)$ small or zero for all θ. If $\mu'(u,\theta) \equiv \frac{\partial}{\partial u} \mu(u,\theta)$ and $\sigma'(u,\theta) = \frac{\partial}{\partial u} \sigma(u,\theta)$ exist, then $h'(u,\theta,\alpha,n) = \frac{\partial}{\partial u} h(u,\theta,\alpha,n)$ gives the Taylor series approximation

$$h(u,\theta,\alpha,n) \doteq h(\theta,\theta,\alpha,n) + h'(\theta,\theta,\alpha,n)\ (u - \theta)$$

and a corresponding alternative form of the above approximation to $a(u,\theta,\theta_n(\cdot,\alpha))$. In the special case of locally-best score quasistatistics, since $\mu(\theta,\theta) \equiv 0$ and $\mu'(\theta,\theta) = \sigma^2(\theta,\theta)$, we find

$$h(u,\theta,\alpha,n) \triangleq \sqrt{n}\sigma(\theta,\theta)(\theta - u) + \Phi^{-1}(\alpha)\left[1 + \frac{\sigma'(\theta,\theta)}{\sigma(\theta,\theta)}(\theta - u)\right] .$$

In the first term, the coefficient $\sqrt{n}\sigma(\theta,\theta)$ of the error $(\theta - u)$ is $\sqrt{I(\theta)}$, where $I(\theta)$ is Fisher's "Information in X at θ." The second term is zero for $\alpha = \frac{1}{2}$ and for the maximum likelihood estimator; for other estimators, the first term dominates the second as n increases. The indicated approximations to risk curves are

$$a(u,\theta,\hat{\theta}_n) \triangleq a(u,\theta,\theta_n(\cdot,.5)) \triangleq \Phi(-\sqrt{n}\sigma(\theta,\theta)\cdot|u - \theta|),$$

and for $\alpha \neq \frac{1}{2}$

$$a(u,\theta,\theta_n(\cdot,\alpha)) \triangleq \begin{cases} \Phi(-\sqrt{n}\sigma(\theta,\theta)(\theta-u) + \Phi^{-1}(\alpha)[\frac{\sigma'(\theta,\theta)}{\sigma(\theta,\theta)}(\theta-u)+1]), & u < \theta \\ 1 - \Phi\ (\ldots \text{same argument} \ldots), & u > \theta, \end{cases}$$

$$\triangleq \text{(more roughly) } \Phi(-\sqrt{n}\sigma(\theta,\theta)\cdot|u - \theta|).$$

These approximations exhibit the approximate normality of distribution of these estimators for large n. While locally best estimators are in general not comparable with other estimators (e.g. those above with $\theta_1(\theta) < \theta_2(\theta)$ for all θ) having similar location functions except in problems of a simple structure, the designation "Information" for $I(\theta)$ is clearly appropriate and useful for cases in which so much precision is attainable that interest is practically restricted to very small $|u - \theta|$, in which case an appropriate choice of an estimator will usually be one which is locally best or perhaps one defined as above with $\theta_2(\theta)-\theta_1(\theta)$ small for all θ.

It should be noted that the preceding approximations which utilize a Taylor series approximation are not accompanied by bounds on errors of approximations. Even in cases where such approximations are very close, under a severly nonlinear transformation of the parameter space ($\theta \longrightarrow \eta = \eta(\theta)$ with $\eta(\theta)$ differentiable and increasing such approximations can become very inaccurate. Hence the principal concrete value of such approximation formulae seems to be that they provide convenient quantitative conjectures which are more or less plausible but which require independent confirmation (or disconfirmation) for specific problems and sample sizes. Similar remarks apply to the preceding approximation formulae based on the Central Limit Theorem only, with the qualification that such approximations could be termed "less asymptotic" than those which also use the Taylor series approximation, in the sense that the former approximations are unaffected by monotone transformations of the parameter space, and their use can be accompanied by use of the known bounds on errors in the Central Limit Theorem approximation.

8.3 Remarks on asymptotic efficiency of estimators.

The theory of the asymptotic efficiency of maximum likelihood estimators (cf. for example Cramer [13], pp. 500-504) utilizes a criterion of asymptotic efficiency (l.c. 489-490) which is restrictive in that it applies only to estimators having asymptotically normal distributions with means equal to the parameter estimated; such estimators are clearly asymptotically median-unbiased (probability of underestimation approaches $\frac{1}{2}$ as n increases). It is advantageous to use a less restrictive criterion of asymptotic efficiency, one which applies to all (sequences of) estimators which are asymptot-

ically median-unbiased. In order to embrace confidence limit estimation as well as point estimation, it is advantageous to define a criterion of asymptotic efficiency which can be applied to any sequence of estimators whose probabilities of underestimation (at each θ) converge with increasing n to a fixed constant α, $0 < \alpha < 1$; any such sequence may be termed an asymptotically valid sequence of confidence limit estimators (of specified coefficient α).

Under broad conditions (some simple ones were given above) consistent estimators exist; it is then natural to define asymptotic efficiency of estimators in terms of the properties of risk curves of estimators in the neighborhood of the true value of θ: an asymptotically efficient sequence of confidence limit estimators may be defined informally as one which is asymptotically valid and asymptotically locally best. The estimators defined above and illustrated in the following section based upon quasistatistics of the form $v_n(x_n,\theta,\alpha) = S(x_n,\theta) - G_n(\theta,\alpha)$ provide examples of such estimators, and have the further properties of being exactly (non-assymptotically) valid and locally-best (and typically admissible). Additional examples are based on quasistatistics of the form $v_n(x_n,\theta,\alpha) = S(x_n,\theta_{1,n}(\theta),\theta_{2,n}(\theta)) - G_n(\theta,\alpha)$ where as n increases $\theta_{2,n}(\theta) - \theta_{1,n}(\theta)$ decreases to zero rapidly enough to give the asymptotically locally-best property; such estimators have the further properties of exact validity and admissibility, and the functions $\theta_{i,n}(\theta)$ can be chosen so that for any finite sample size a suitable emphasis is given to avoiding errors exceeding specified positive magnitudes; for practical applications, such estimators seem preferable in principle to (exactly) locally-best estimators.

The usual asymptotic theory (l.c.) is free of the important assumption (b) of Corollary 1 above. From the present non-asymptotic standpoint, for each θ the acceptance region $A(\theta) = \{x | S(x,\theta) \leqq 0\}$ represents a locally-best one-sided test, and the family of such tests can be used as usual to define a confidence region for estimation of θ, namely $U(x) = \{\theta | x \in A(\theta) \theta \in \Omega\}$; in general such a confidence region will not have a constant confidence coefficient, but its theory and interpretation in applications follow usual lines. The failure of assumption (b) corresponds to the failure of the sets $A(\theta)$ to constitute a nondecreasing sequence in θ; this in turn corresponds to the fact that, for some x, the confidence region $U(x)$ will fail to constitute an interval $[\theta^*(x), \overline{\theta}]$ which can be described by a lower estimator $\theta^*(x)$. The theory of admissible confidence regions not necessarily of interval form, and their interpretation in applications, lie outside the scope of the present paper. However, from the present standpoint it may be observed that the principal role of the regularity assumptions in, for example, Cramer (l.c.) is to guarantee that with increasing n, for each θ the probability that $U(x)$ will be an interval (or equivalently that $S(x_n,\theta)$ will satisfy the assumptions of Corollary 1) approaches unity: More precisely with increasing n, for each θ the probability of the set of points x_n on which $S(x_n,u)$ is decreasing in u (at least for u near θ) approaches unity. The key step of the derivation from this standpoint is the observation that the law of large numbers applies, when θ is true, to the sum $\frac{\partial}{\partial u} S(X_n,u) = \sum_{i=1}^{n} \frac{\partial}{\partial u} S(Y_i,u)$, each term of which has (at least for u near θ) a negative expected value $E[\frac{\partial^2}{\partial u^2} \log f(Y_1,u) | \theta]$. (Similar remarks apply to use of generalized score quasistatistics which fail

to satisfy condition (b) of Corollary 1.) Dropping the qualification "for u near θ" gives that the probability of multiple roots of $S(x_n,\theta) = 0$ approaches zero with increasing n. Asymptotic efficiency properties of confidence limits and intervals defined by use of quasistatistics of the form $S(x_n,\theta) - G_n(\theta,\alpha)$ were proved under broad regularity conditions by Wald [14].

The remarks of Lehmann [15], on the limited value of any exclusively-asymptotic theory of optimum tests apply with equal force to estimation theory. Asymptotically efficient estimators may approach efficiency at arbitrarily slow rates as n increases. Only on the basis of an auxiliary non-asymptotic investigation of the quantitative and/or qualitative (optimality) properties of an asymptotically efficient estimator can it be recommended in an application with a specified (finite) sample size.

9. Examples. Examples 1-3 illustrate that the formal treatment of Section 8 can often be applied conveniently to problems admitting uniformly best estimators.

Example 1. Normal mean. Let $x = (y_1,\dots y_n)$ be a sample of n independent observations from a normal distribution with known variance, say $\sigma^2 = 1$, and unknown mean θ, $-\infty < \theta < \infty$. Then

$$f(x,\theta) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2}.$$

Let

$$v(x,\theta) = \frac{\partial \log f(x,\theta)}{\partial\theta} - G(\theta,a(\theta)),$$

where $\alpha(\theta)$ is a given function. Then

$$v(x,\theta) = n(\bar{y} - \theta) - G(\theta,\alpha(\theta)) = n\bar{y} - n\theta - \sqrt{n}\,\Phi^{-1}(\alpha(\theta)) \ ,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\Phi(u)$ is the standard normal c.d.f. Then $v(x,\theta)$ clearly satisfies the conditions of Corollary 1 if $\alpha(\theta)$ is such that $\theta + \frac{1}{\sqrt{n}}\Phi^{-1}(\alpha(\theta))$ is increasing in θ; as n increases, the latter condition becomes a less restrictive one on $\alpha(\theta)$; it is obviously satisfied if $\alpha(\theta) \equiv \alpha$, $0 \leq \alpha \leq 1$. For each such function $\alpha(\theta)$, an admissible estimator $\theta^*(x)$ is defined as the solution θ of $v(x,\theta) = 0$, that is, of

$$\theta + \frac{1}{\sqrt{n}}\Phi^{-1}(\alpha(\theta)) = \bar{y} \ .$$

Denoting the solution by $Q(\bar{y})$, this gives $\theta^*(x) = Q(\bar{y})$; $Q(\bar{y})$ can be any increasing function of $\bar{y}$ if $\alpha(\theta)$ is suitably chosen. For $\alpha(\theta) = \alpha$, this becomes (in the general case where σ^2 is any positive number)

$$\theta^*(x) = \theta(x,\alpha) = \bar{y} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\alpha) \ ,$$

an upper confidence limit of confidence coefficient $1-\alpha$ (and/or a lower confidence limit of coefficient α). Each of these estimators is, by Lemma 2 above, uniformly best among all estimators with the same location functions $a(\theta -,\theta) \equiv 1-a(\theta +,\theta) \equiv \alpha(\theta)$. Taking $\alpha(\theta) \equiv \frac{1}{2}$ gives $\hat{\theta}(x,.5) = \theta(x) = \bar{y}$. Since this estimator is

independent of the value assumed for σ^2, the classical (maximum likelihood and mean-unbiased) estimator $\bar{y}$ is uniformly best among all median-unbiased estimators of θ even if σ^2 is not known. The same property clearly holds for the classical least squares estimators of linear regression theory under normality assumptions.

Example 2. Normal variance. Let $x = (y_1, \ldots y_n)$ be a sample of n independent observations from a normal distribution with known mean, say $\mu = 0$, and unknown standard deviation $\theta = \sigma$, $0 < \sigma < \infty$. Then

$$f(x,\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2} .$$

Let $v(x,\theta) = \frac{\partial}{\partial\theta} \log f(x,\theta) - G(\theta,a(\theta))$, where $a(\theta)$ is a given function. Then

$$v(x,\theta) = \frac{n}{\sigma}\left(\frac{s^2}{\sigma^2} - 1\right) - G(\sigma,a(\sigma)) ,$$

where $s^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$ is the usual unbiased estimator of σ^2. For a given σ, $\frac{ns^2}{\sigma^2}$ has the Chi-Square distribution with n degrees of freedom; hence $G(\sigma,a(\sigma)) = \frac{1}{\sigma}(\chi^2_{n,a(\sigma)} - n)$, where $\chi^2_{n,a}$ is the lower a-point of the Chi-square distribution with n degrees of freedom. Thus $v(x,\theta) = \frac{1}{\sigma}\left(\frac{ns^2}{\sigma^2} - \chi^2_{n,a(\sigma)}\right)$. If, for example, $a(\sigma) = a$, then

$$\theta^*(x) = \theta(x,a) = s\sqrt{n/\chi^2_{n,a}}$$

which is a uniformly best estimator, by Lemma 2 above. A uniformly

best median-unbiased estimator of σ is $\sigma(x,.5)$. Similarly, uniformly best estimators of the variance σ^2 are given by

$$\sigma^2(x,\alpha) = s^2 n/\chi^2_{n,\alpha} \quad .$$

When n is not small, $n/\chi^2_{n,.5} \doteq 1$, and $\sigma(x,.5) \doteq s$ and $\sigma^2(x,.5) \doteq s^2$. Thus the commonly used point estimators s and s^2 can be justified on the grounds that they are uniformly best (among estimators with the same location functions) and very nearly (except when n is very small) median-unbiased. Tables of the Chi-square distribution provide the constants $\chi^2_{n,.5}$, which can be used in place of n in standard procedures for computing s or s^2, to obtain the estimates $\sigma(x,.5)$ or $\sigma^2(x,.5)$ respectively. Comparisons of these and other estimators from the standpoint of median-bias, with tables, were given by Eisenhart and Martin [16]. For the more usual problem in which μ is unknown, with N = n+1 observations, the same remarks apply to the usual mean-unbiased estimator $s^2 = \sum_{i=1}^{N} (y_i - \bar{y})/(N - 1)$ and to s. The theory of such multi-parameter problems lies outside the formal scope of the present paper.

Example 3. Binomial mean. Let $x = (y_1, \ldots y_n)$, where the Y_i's are independent, Prob $(Y_i = 1) = \theta$, Prob $(Y_i = 0) = 1 - \theta$, $0 \leq \theta \leq 1$. Let Z be an auxiliary randomization variable, uniformly distributed on $0 \leq Z < 1$. Then $t = t(x,z) = n\bar{y} + z$, where $n\bar{y} = \sum_{i=1}^{n} y_i$ is a sufficient statistic having the monotone likelihood ratio property; hence each nondecreasing function $\theta^*(t)$ taking values in the unit interval is a uniformly best estimator. The classical (maximum likelihood, unbiased) estimator is $\theta = [t]/n = \bar{y}$, where [t]

is the largest integer not exceeding t. By use of binomial tables, exact confidence limits $\theta(t,\alpha)$ and median unbiased estimators $\theta(t,.5)$ can be determined easily as the solutions θ of the equations $\alpha = \text{Prob}\,(T \leqq t|\theta)$, where t is the observed value of the statistic. For typical purposes of informative inference, it seems preferable to dispense with use of the randomization variable z; a non-randomized uniformly best point estimator having location functions closest to $\frac{1}{2}$, in a certain sense, is defined for each observed value of $n\bar{y}$ as the solution θ of the equation $\text{Prob}\,(Y < \bar{y}|\theta) = \text{Prob}\,(\bar{Y} > \bar{y}|\theta)$; this estimator $\hat{\theta}(\bar{y})$ is easily determined by use of binomial tables; when n is not small, we have $\hat{\theta}(\bar{y}) \doteq \bar{y}$. In all cases the effect of the randomization variable is minor except when n is small. Thus the classical mean-unbiased estimator can be justified on the grounds that it is uniformly best (among estimators with the same location functions) and is very nearly (except when $n\theta$ or $n(1-\theta)$ is very small) median-unbiased

Other discrete examples with the m.l.r. property, such as the Poisson and negative binomial, may be treated similarly.

Example 4. Logistic mean. Let $x = (y_1,\ldots y_n)$ be a sample of n independent observations from a logistic distribution with unknown mean θ: $\text{Prob}\,(Y \leqq y|\theta) = \Psi(y-\theta) = (1+e^{-(y-\theta)})^{-1}$, $-\infty < y < \infty$, $-\infty < \theta < \infty$; Y has the density function

$$\psi(y-\theta) = e^{-(y-\theta)}/(1+e^{-(y-\theta)})^2, \quad -\infty < y < \infty .$$

For any fixed $\Delta > 0$, taking $\theta_1(\theta) = \theta - \Delta$, $\theta_2(\theta) = \theta + \Delta$, determines a score quasistatistic

$$S(x,\theta-\Delta,\theta+\Delta) = \frac{1}{2\Delta}\left[\sum_{i=1}^{n} (\log \psi(y_i-\theta-\Delta) - \log \psi(y_i-\theta+\Delta)\right].$$

For any fixed α, $0 \leqq \alpha \leqq 1$, taking $\alpha(\theta) \equiv \alpha$ determines a score quasistatistic

$$v(x,\theta,\alpha) = S(x,\theta-\Delta,\theta+\Delta) - G(\theta,\alpha)$$

which satisfies the conditions of Corollary 1 of Section 6 above, and hence determines an admissible confidence limit estimator $\theta^* = \theta(x,\alpha)$ as the solution θ of the equation $v(x,\theta,\alpha) = 0$. Since θ is translation parameter, $G(\theta,\alpha)$ is independent of θ, and may be written $G(\alpha)$. By symmetry, $G(.5) = 0$. $G(\alpha)$ can be determined approximately, except for α very near 0 or 1 and for very small n, by use of the Central Limit Theorem: let $\mu(u,\theta)$ and $\sigma^2(u,\theta)$ denote respectively the mean and variance of $S(Y,u-\Delta,u+\Delta)$ when θ is true; then $\mu(\theta,\theta) = 0$ by symmetry; we may write $\mu(u-\theta)$ and $\sigma^2(u-\theta)$ because θ is a translation parameter. We have

$$\text{Prob}\left\{v(X,u,\alpha) \leqq 0 \mid \theta\right\} \doteq \Phi\left(\frac{G(\alpha)-n\mu(u-\theta)}{\sqrt{n}\,\sigma(u-\theta)}\right)$$

which provides an approximation to the risk curves $a(u,\theta,\theta^*)$ of the estimator $\theta^* = \theta(x,\alpha)$; for the determination of $G(\alpha)$, similarly

$$\text{Prob}\left\{v(X,\theta,\alpha) \leqq 0 \mid \theta\right\} \equiv \alpha \doteq \Phi(G(\alpha)/\sqrt{n}\,\sigma(0)), \text{ or } G(\alpha) \doteq \sqrt{n}\sigma(0)\Phi^{-1}(\alpha).$$

This, with the formula above gives the approximate risk curves of θ^*

$$a(u,\theta,\theta^*) \doteq \begin{cases} \Phi\left(- \sqrt{n}\,\frac{\mu(u-\theta)}{\sigma(u-\theta)} + \frac{\sigma(0)}{\sigma(u-\theta)}\,\Phi^{-1}(\alpha)\right) & \text{for } u < \theta. \\ 1 - \Phi(\text{...same argument...}) & \text{for } u > \theta. \end{cases}$$

The preceding discussion depended throughout on the chosen value $\Delta > 0$. A locally best confidence limit estimator $\theta^* = \theta(x,\alpha)$ is determined as the solution θ of the equation

$$v(x,\theta,\alpha) \equiv S(X,\theta) - G(\alpha) = 0.$$

Here $S(y,\theta) = \frac{\partial}{\partial\theta} \log \Psi(y-\theta) = 2\Psi(y-\theta) - 1$; $\Psi(Y-\theta)$ has, when θ is true, a uniform distribution on the unit interval; hence when θ is true the c.d.f. of $\sum_{i=1}^{n} \Psi(Y_i-\theta)$ (and hence that of $S(X,\theta)$) can be calculated as in Cramer [13], pp. 244-246. The normal approximation gives (since

$$\sigma^2(0) = \mathrm{Var}[S(Y,\theta)|\theta] = \tfrac{1}{3},\ \mathrm{Var}\,[S(X,\theta)|\theta] = \tfrac{n}{3}),\ G(\alpha) \doteq \sqrt{\tfrac{n}{3}}\,\Phi^{-1}(\alpha);$$

$\alpha = \frac{1}{2}$ gives exactly $G(\frac{1}{2}) = 0$ and determines the maximum likelihood estimator $\hat{\theta} = \theta(x,.5)$. In general, a locally best confidence limit estimator $\theta(x,\alpha)$ is determined (approximately, except for $\alpha = \frac{1}{2}$) as the root θ of the equation $S(x,\theta) = \sqrt{\frac{n}{3}}\,\Phi^{-1}(\alpha)$, or

$$\sum_{i=1}^{n} \Psi(y_i-\theta) \doteq \frac{n}{2} + \frac{1}{2}\sqrt{\frac{n}{3}}\,\Phi^{-1}(\alpha)\ .$$

Such an equation is easily solved numerically by use of Berkson's tables of $\Psi(u)$ ([17]).

The present example serves also to illustrate the determination of an admissible confidence curve estimator by use of a family of quasistatistics as described at the end of Section 6 above. Each of the families of quasistatistics $v(x,\theta,\alpha)$, $0 \leqq \alpha \leqq 1$ considered here (each based upon a fixed $\Delta \geqq 0$) has the property that $\theta(x,\alpha)$ is, for each fixed x, decreasing in α; in fact, for each x, $\theta(x,\alpha)$ decreases continuously from ∞ to $-\infty$ as α increases from 0 to 1. Thus for each observed x, each θ ($-\infty \leqq \theta \leqq \infty$) will be a confidence limit $\theta(x,\alpha)$ for some α; we can conveniently determine the required solutions $\theta(x,\alpha)$ of $v(x,\theta,\alpha) = 0$ in the form

$$\alpha(x,\theta) = \text{Prob}\left\{S(X,\theta) \leqq S(x,\theta) \,|\, \theta\right\} \doteq \Phi\left(\sqrt{\tfrac{3}{n}}\, S(x,\theta)\right)$$

for as many values of θ as desired.

Numerical example. Let $x = (y_1,y_2,y_3) = (0,0,6)$. Letting θ_i denote a trial value of θ, $S_i = S(x,\theta_i)$, and $\alpha_i = \alpha(x,\theta_i)$, $\text{Prob}\left\{S(X,\theta_i) \leqq S(x,\theta_i) \,|\, \theta_i\right\}$, $i = 1,2,\ldots,$ and taking $\theta_1 = \bar{y} = 2$ as a trial value plausibly near $\theta(x,.5) = \hat{\theta}$, we obtain

$$S_o = 2\sum_{i=1}^{3}\Psi(y_i-2) - 3 = -0.559, \quad \alpha_1 \doteq \Phi(-.559) = .288.$$

Further similar computations are summarized in Table 1 and in Fig. 7 a sketch of the confidence curve $c(\theta,x) = \min\,[\alpha(x,\theta),\, 1-\alpha(x,\theta)]$.

Table I

i	θ_i	S_i	approx. a_i	exact a_i
1	2.0	-0.559	.288	
2	1.44	-0.256	.399	
3	1.18	-0.758	.470	
4	1.12	-0.031	.488	
5	1.08	-0.0005	.4998	.4998
6	3.08	-0.927	.177	
7	4.0	-1.166	.122	
8	5.0	-1.511	.065	
9	6.0	-2.0	.023	
10	7.0	-2.462	.007	
11	-1.0	1.924	.973	
12	-2.0	2.523	.994	.998
13	0.0	1.0	.841	.833

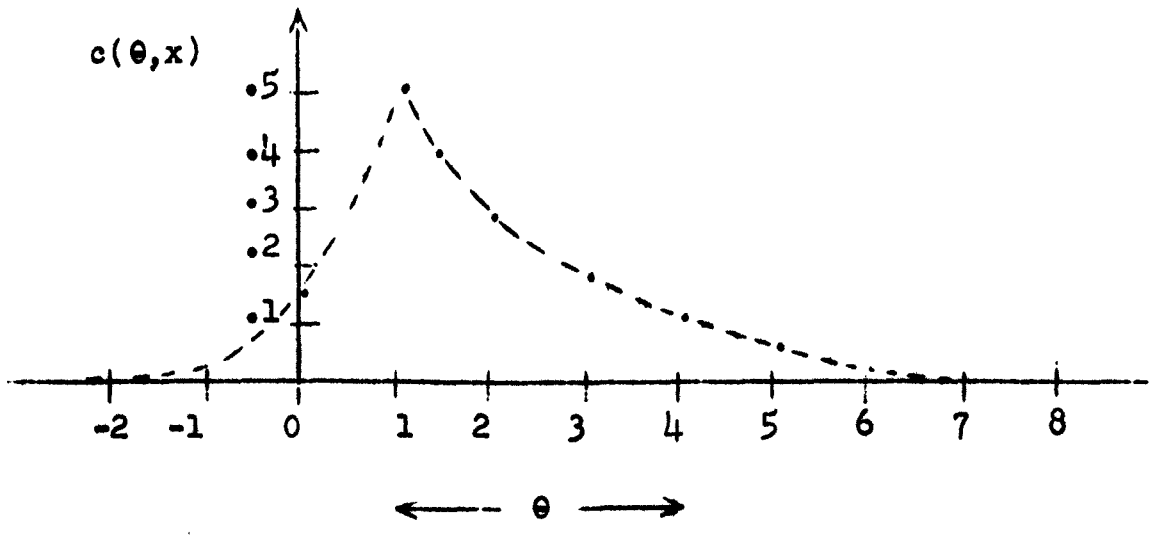


Figure 1

The closeness of the normal approximations can be checked in the present case by use of the exact formula (based on Cramer, l.c.

$$\alpha(x,\theta) = \begin{cases} \frac{z^3}{6} & 0 \leqq z \leqq 1 , \\ \frac{z^3}{6} - \frac{1}{2}(z-1)^3 , & 1 \leqq z \leqq 2 , \\ 1 - \frac{1}{6}(3-z)^3 , & 2 \leqq z \leqq 3 , \end{cases}$$

where $z = z(x,\theta) = \frac{1}{2}(S(x,\theta) + 3)$. The approximation is seen to be quite adequate here. In other examples, if exact values of $\alpha(x,\theta)$ cannot be obtained by use of standard tables or tractable integrals, one may consider checking approximate values of $\alpha(x,\theta)$, for a few values of θ of particular interest, by use of (a) the error-bound on the normal approximation, (b) numerical integration, (c) empirical sampling (Monte Carlo), or possibly (d) an asymptotic expansion. For (a) and (d), see Wallace, [18].

The values θ_i above, for $i = 2,\ldots 5$, were determined by $\theta_{i+1} = \theta_i + S_i$, based on Fisher's formula $\theta_{i+1} = \theta_i + S(x,\theta_i)/\mathrm{Var}\,[S(X,\theta_i)|\theta_i]$ for iterative calculation of maximum likelihood estimates. If $\log f(x,\theta) \doteq a\theta^2 + b\theta + c$ for some constants $a < 0, b, c$, at least for θ near $\hat{\theta}(x)$ (asymptotic theory shows that this will be the case with high probability for sufficiently large n, under certain regularity conditions), then $S(x,\theta) \doteq 2a\theta + b, \frac{\partial}{\partial\theta} S(x,\theta) = 2a$; $(a\theta^2 + b\theta + c)$ is minimized by $\theta^* = -b/2a \doteq \theta - S(x,\theta)/\frac{\partial}{\partial\theta} S(x,\theta)$. $\frac{\partial}{\partial\theta}S(x,\theta)$ may be calculated directly; or approximated numerically from difference quotients $\frac{\Delta S(x,\theta)}{\Delta\theta}$ based on previously calculated θ_i's; or (as done above)

"estimated" by its expected value: for sufficiently large n, with high probability the approximation

$$\frac{\partial}{\partial\theta}S(x,\theta) \doteq E[\frac{\partial}{\partial\theta}S(X,\theta)|\theta] = E[\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)|\theta]$$

$$= -\text{Var}\,[S(X,\theta)|\theta] = -I(\theta)$$

is effectively close. The rate of convergence of θ_i to $\hat{\theta}$ may be slow as above, for samples with "improbable configurations" and/or small n; use of $\frac{\partial}{\partial\theta}S(x,\theta)$ rather than its expected value here would evidently give faster convergence, but would require additional calculations for each i. Speed of convergence is not of exclusive interest here; since a number of values of $\alpha_i = \alpha(x,\theta_i)$ are desired for a sketch of the confidence curve estimate, any convenie method of choosing successive θ_i's may be used.

The values θ_6 and θ_{11} above were chosen as trial approximations to the confidence limits $\theta(x,.025)$, $\theta(x,.975)$ respectively, by use of the asymptotic formula for such confidence limits:

$$\hat{\theta} \pm \mathcal{I}^{-1}(.975)/\text{Var}\,[S(X,\theta)|\hat{\theta}] \doteq \hat{\theta} \pm 2\,.$$

The poor approximations obtained provide a limited illustration of the fact that such approximations are "more asymptotic," i.e. may be expected to be often less close, than the normal approximations to distributions of score statistics.

Example 5. Laplacean mean. Let $x = (y_1,\ldots y_n)$ be a sample of n independent observations from a Laplacean (double exponential) distribution with unknown mean θ, $-\infty < \theta < \infty$, with density

function

$$h(y,\theta) = \tfrac{1}{2} e^{-|y-\theta|}, \quad -\infty < y < \infty \ .$$

For any fixed $\Delta > 0$, let $v(x,\theta,\alpha) = S(x,\theta - \Delta, \theta + \Delta) - G(\theta,\alpha)$

$$= \frac{1}{2\Delta}\left[\sum_{i=1}^{n} (|y_i - \theta - \Delta| - |y_i - \theta + \Delta|)\right] - G(\alpha)$$

We note that

$$|y - \theta - \Delta| - |y - \theta + \Delta| = \begin{cases} 2\Delta & \text{if } \theta \leqq y - \Delta \\ 2(y - \theta) & \text{if } y - \Delta \leqq \theta \leqq y + \Delta \\ -2\Delta & \text{if } y + \Delta \leqq \theta, \end{cases}$$

and hence

$$-2\Delta n \leqq \sum_{i=1}^{n} (|y_i - \theta - \Delta| - |y_i - \theta + \Delta|) \leqq 2\Delta n \text{ for all } x.$$

Since $\text{Prob}\left\{Y \leqq \theta - \Delta | \theta\right\} = \tfrac{1}{2} e^{-\Delta}$, the c.d.f. of $\sum_{i=1}^{n}(|Y_i - \theta - \Delta| - |Y_i - \theta + \Delta|)$ has a jump of $(\tfrac{1}{2} e^{-\Delta})^n$ at each end of its range, and is continuously increasing between these jumps. Hence $G(\alpha)$ is well-defined if $(\tfrac{1}{2}e^{-\Delta})^n < \alpha < 1 - (\tfrac{1}{2}e^{-\Delta})^n$; for other α's use of an auxiliary randomization variable would be necessary; by symmetry, $G(\tfrac{1}{2}) = 0$. A simple computation gives

$$\text{Var}\,(|Y - \theta - \Delta| - |Y - \theta + \Delta|) = 8(1 - e^{-\Delta} - \Delta e^{-\Delta}), = v,$$

say; for n not very small and α not extreme, the normal approximation to the distribution of $v(X,\theta,\alpha)$ gives

$$G(\alpha) \doteq \sqrt{nv}\, \Phi^{-1}(\alpha) \quad .$$

For any α bounded as above, by Corollary 1 the estimator $\Theta(x,\alpha)$, defined as the solution θ of $v(x,\theta,\alpha) = 0$, is admissible.

The median-unbiased estimator $\Theta(x) = \Theta(x,.5)$ defined as the solution θ of

$$\sum_{i=1}^{n} |(y_i - \Delta) - \theta| = \sum_{i=1}^{n} |(y_i + \Delta) - \theta|$$

(which is easily solved numerically), depends upon the particular value Δ chosen; the error-probabilities $a(\theta - \Delta,\theta,\theta^*)$, $a(\theta + \Delta,\theta,\theta^*)$ have a minimized common value for all θ.

Locally-best estimators ("$\Delta \longrightarrow 0$") $\Theta(x,\alpha)$ are defined by use of

$$v(x,\theta,\alpha) = \sum_{i=1}^{n} I(y_i > \theta) - \sum_{i=1}^{n} I(y_i < \theta) - G(\alpha) \ ,$$

where, for any relation R, the indicator-function I(R) is defined by I(R) = 1 if R is true and I(R) = 0 if R is false. Thus $\sum_{i=1}^{n} I(y_i > \theta) - \sum_{i=1}^{n} I(y_i < \theta)$ is the number of observations y_i exceeding θ minus the number of observations less than θ; with probability one, the observations y_i have n distinct values, and may be ordered, $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$. Then

$$\sum_i I(y_i > \theta) - \sum_i I(y_i < \theta) = \begin{cases} n, & \text{if } \theta < y_{(1)}, \\ n-1, & \text{if } \theta = y_{(1)} \\ n-2, & \text{if } y_{(1)} < \theta < y_{(2)}, \\ \vdots & \\ -n+1, & \text{if } \theta = y_{(n)}, \text{ and} \\ -n, & \text{if } \theta > y_{(n)}. \end{cases}$$

Let r be any integer, $1 \leq r \leq n$. It is easily seen that for

$$\alpha \equiv 1 - \frac{1}{2^n} \sum_{u=0}^{n-r} \binom{n}{r}, \quad G(\alpha) \equiv n + 1 - 2r;$$

hence

$$v(x,\theta,\alpha) = \sum_{i=1}^{n} I(y_i > \theta) - \sum_{i=1}^{n} I(y_i < \theta) - (n + 1 - 2r) .$$

With probability one, $v(x,\theta,\alpha) = 0$ will have a unique solution, namely $\theta(x,\alpha) = y_{(r)}$. Since $G(0) = -n$ and $G(1) = n$, $\theta(x,1) \equiv -\infty$ and $\theta(x,0) \equiv \infty$. For any observed x, the set of (n+2) confidence limits

$$[\theta(x,1),\theta(x,1-(\tfrac{1}{2})^n),..\theta(x,(\tfrac{1}{2})^n),\theta(x,0)] \equiv [-\infty,y_{(1)},y_{(2)}..y_{(n)},\infty]$$

serves as a (locally-best) confidence curve estimate. (For other values of α, use of an auxiliary randomization variable would be required in defining $v(x,\theta,\alpha)$.) In contrast to the approximate confidence limits given by asymptotic methods, the various exact confidence limits here depend on all values y_i in the sample x and not only on the value of $\hat{\theta} = y_{((n+1)/2)}$, the sample median (for n odd).

For the more general problem of estimating the median θ of a Laplacean density function

$$h(y,\theta,c) = \frac{1}{2c} e^{-|y-\theta|/c}, \quad -\infty < y < \infty ,$$

with known scale parameter $c > 0$, similar derivations give the same locally best confidence limits and confidence curve estimators. Since these estimators are independent of c, they can be used for estimation of θ in the more general problem in which c is unknown. For the latter problem, they remain valid and locally best (with respect to errors in estimation of θ, uniformly in c), and their risk curves respectively depend on the argument $(u-\theta)/c$.

Still more generally, let the Y_i's be independent with any continuous c.d.f. of unknown form, with unknown median θ. Since the estimators of θ given above remain valid (have the given location functions), and are essentially unique locally-best estimators with the given location functions in the special case of Laplacean distributions, these estimators may be called admissible for the non-parametric problem of estimation of a median of a (continuous) distribution of unknown form. Similar remarks apply to such use of order statistics $y_{(i)}$ as estimators of the p-quantile of a continuous distribution of unknown form; here the generalized Laplacean density function

$$h(y,\theta) = \begin{cases} pe^{-|y-\theta|}, & y < \theta , \\ (1-p)e^{-|y-\theta|}, & y \geq \theta , \end{cases}$$

for which θ is the p-quantile, replaces the Laplacean density, for any specified p, $0 < p < 1$, and the derivation proceeds in essentially the same way as above where $p = \frac{1}{2}$.

Example 6. Quantal response models. Let $x = (y_1, \ldots y_n)$, where the Y_i's are independent,

$$\text{Prob}\left\{Y_i = 1|\theta\right\} = P_i(\theta), \quad \text{Prob}\left\{Y_i = 0|\theta\right\} = Q_i(\theta) = 1 - P_i(\theta),$$

$$i = 1, \ldots n,$$

where the $P_i(\theta)$'s are known increasing functions of θ, having derivatives $P_i'(\theta)$, $\theta \in \Omega = (\underline{\theta}, \overline{\theta})$, an open interval. Examples include: (1) Dilution series [19]: $P_i(\theta) = 1 - e^{-d_i\theta}$, where d_i is a known "dose" (volume) of material examined in the i^{th} observation, and θ is the unknown mean concentration of minute particles per unit volume randomly distributed in the material. (2) Mental ability tests, normal model [20]: $P_i(\theta) = (1/k_i) + ((k_i-1)/k_i)\Phi(a_i+b_i\theta)$ is the probability that a subject with unknown ability-parameter θ will respond correctly to the i^{th} item in a test. Here Φ is the standard normal c.d.f., and the parameters $0 < k_i \leqq \infty$, $-\infty < a_i < \infty$, and $b_i > 0$ which characterize the i^{th} item may be assumed known (or estimated with high precision) on the basis of previous investigation; a_i represents the item's level of difficulty, b_i its sensitivity, and $(1/k_i)$ if positive may be interpreted as the probability of a correct response due to guessing only. (3) Mental ability test, logistic model [21]: As in (2), with $\Phi(u)$ replaced by the logistic c.d.f. $\Psi((1.7)u) = 1/(1 + e^{(-1.7)u})$. This very slight quantitative modification gives a model which is equally plausible and has much

greater mathematical tractability; in the case where $1/k_i = 0$, it provides a sufficient statistic with the monotone likelihood ratio property. (4) One-parameter bioassay model, normal form [22]: $P_i(\theta) = (1/k) + ((k-1)/k)\Phi(\theta + bd_i)$. Here θ is the unknown concentration of a component in material being assayed; the case $1/k = 0$ is most common; d_i is a known dose parameter; b is a sensitivity parameter which in special cases may be known or estimated with relatively high precision. (5) One-parameter bioassay model, logistic form [23]: As in (4), with Φ replaced by Ψ. In the usual case $1/k = 0$, with b known this model provides a sufficient statistic with the monotone likelihood ratio property.

We have

$$S(y_i,\theta) = \begin{cases} P_i'(\theta)/P_i(\theta) & \text{for } y_i = 1, \\ Q_i'(\theta)/Q_i(\theta) = -P_i'(\theta)/(1-P_i(\theta)) & \text{for } y_i = 0, \end{cases}$$

or

$$S(y_i,\theta) = P_i'(\theta)/Q_i(\theta) + y_i P_i'(\theta)/P_i(\theta)Q_i(\theta), \qquad y_i = 0 \text{ or } 1,$$

and

$$\mu_i(u,\theta) = E[S(Y_i,u)|\theta] = P_i'(u)\left[\frac{P_i(\theta)}{P_i(u)} - \frac{Q_i(\theta)}{Q_i(u)}\right],$$

$$\sigma_i^2(u,\theta) = \text{Var}\,[S(Y_i,u)|\theta] = P_i'(u)^2[P_i(\theta)Q_i(\theta)/P_i(u)^2Q_i(u)^2],$$

$$\mu_i(\theta,\theta) = 0, \sigma_i^2(\theta,\theta) = P_i'(\theta)^2/P_i(\theta)Q_i(\theta).$$

The normal approximation gives

$$\text{Prob}\,[S(X,u) \leqq k|\theta] \doteq \mathbb{\Phi}\left(\frac{k-\sum_i \mu_i(u,\theta)}{\left(\sum_i \sigma_i^2(u,\theta)\right)^{1/2}}\right) .$$

For a given (u,θ), this approximation is close provided that (a) the right member is not very near 0 nor 1, and (b) the number m of $\sigma_i^2(u,\theta)$'s near $\max_i \sigma_i^2(u,\theta)$ in value is not small.

If for each i and y_i, $S(y_i,\theta)$ is decreasing in θ (i.e., $P_i(\theta)P_i''(\theta) < P_i'(\theta)^2$ and $Q_i(\theta)P_i''(\theta) < P_i'(\theta)^2$), then $v(x,\theta) = S(x,\theta)$ satisfies the conditions of Corollary 1, and the maximum likelihood estimator $\hat{\theta}(x)$, the solution of $S(x,\theta) = 0$, is admissible; if the normal approximation above (with $u = \theta$) is close for respective values of θ, θ is approximately median-unbiased; if the approximation is close for respective values of (u,θ), $\hat{\theta}$ has the approximate risk curves

$$a(u,\theta,\hat{\theta}) \doteq \begin{cases} \mathbb{\Phi}(-\sum_i \mu_i(u,\theta)/(\sum_i \sigma_i^2(u,\theta))^{1/2}), & u < \theta , \\ 1-\mathbb{\Phi}(\ldots \text{same argument} \ldots), & u > \theta . \end{cases}$$

More generally, to determine locally best (approximate) confidence limits $\theta(x,\alpha)$ as solutions θ of

$$v(x,\theta,\alpha) \equiv S(x,\theta) - (\sum_i \sigma_i^2(\theta,\theta))^{1/2}\mathbb{\Phi}^{-1}(\alpha) = 0,$$

a simple adaptation of the discussion at the end of Section 8.1 above may be applied to the problem of verification of the conditions of Corollary 1.

Example 7. Rectangular mean Let $x = (y_1, \ldots y_n)$ be a sample of n independent observations on a random variable Y with density

$$h(y,\theta) = \begin{cases} 1 \text{ if } \theta - \frac{1}{2} \leq y \leq \theta + \frac{1}{2} \\ 0 \text{ otherwise,} \end{cases}$$

with $\theta = E(Y)$ unknown. Let r and s denote respectively the smallest and the largest of the observed values y_i. Let $\theta^* = \theta^*(r,s)$ be any function, defined for all r,s such that $r \leq s \leq r + 1$, which satisfies $s - \frac{1}{2} \leq \theta^*(r,s) \leq r + \frac{1}{2}$ and which is nondecreasing in r and in s. Then $\theta^*(r,s)$ satisfies the conditions of Lemma 1 since, for each θ_0, $\{x|\theta^* \leq \theta_0\}$ and $\{x|\theta^* < \theta_0\}$ satisfy the (necessary and) sufficient condition given by Pratt [24] for admissibility of one-sided tests on θ. Venketeraman [25] has shown that such estimators constitute an essentially complete class, and has given minimal complete and minimal essentially complete classes of estimators of θ.

For samples of size n = 2, each of the following estimators is admissible and median-unbiased:

$\theta^*(x) = (r + s)/2$, the usual mean-unbiased estimator.

$$\theta'(x) = \begin{cases} s - \frac{1}{2}, & \text{if } s \geq r + 1/\sqrt{2}, \\ r + (\sqrt{2} - 1)/2, & \text{if } s \leq r + 1/\sqrt{2}, \end{cases}$$

$$\theta''_2(x) = \begin{cases} r + \frac{1}{2}, & \text{if } r \leq s - 1/\sqrt{2}, \\ s - (\sqrt{2} - 1)/2, & \text{if } r \geq s - 1/\sqrt{2}. \end{cases}$$

Among median-unbiased admissible estimators, θ' is uniformly best with respect to errors of under-estimation, and θ'' is uniformly best with respect to errors of over-estimation. Analogous confidence curve estimators are easily constructed.

For any fixed k, $0 \leqq k \leqq \frac{1}{2}$, for testing hypotheses of the form $H(\theta_0)$: $\theta \leqq \theta_0$ or $H(\theta_0-)$: $\theta < \theta_0$, there is an admissible acceptance region

$$A(\theta_0) = \left\{x \mid \frac{r+s}{2} \leqq \theta_0 + k, \quad s \leqq \theta_0 + \tfrac{1}{2}\right\}$$

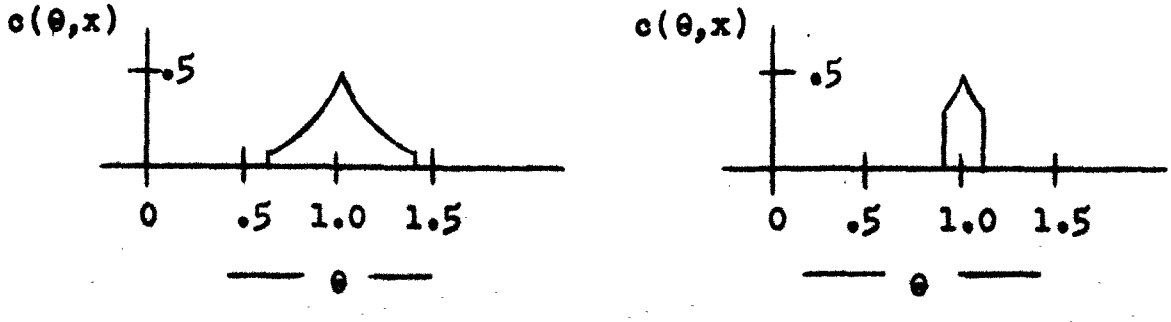
and another admissible acceptance region

$$A'(\theta_0) = \left\{x \mid \frac{r+s}{2} \leqq \theta_0 - k, \text{ or } r \leqq \theta_0 - \tfrac{1}{2}\right\} .$$

From such tests we obtain admissible confidence limit estimators at each level, and the corresponding admissible confidence curve estimator:

$$c(\theta,x) = \begin{cases} 0, \text{ if } \theta \geqq r + \frac{1}{2} \text{ or } \theta \leqq s - \frac{1}{2}, \\ 2[\frac{1}{2} - |\theta - \frac{r+s}{2}|]^2, \text{ otherwise} . \end{cases}$$

If $x = (0.9,1.1) = (r,s)$, or alternatively if $x = (0.6,1.4) = (r,s)$, we obtain respective confidence curve estimates which reflect that the "amount of information in a sample" increases with $(s-r)$:



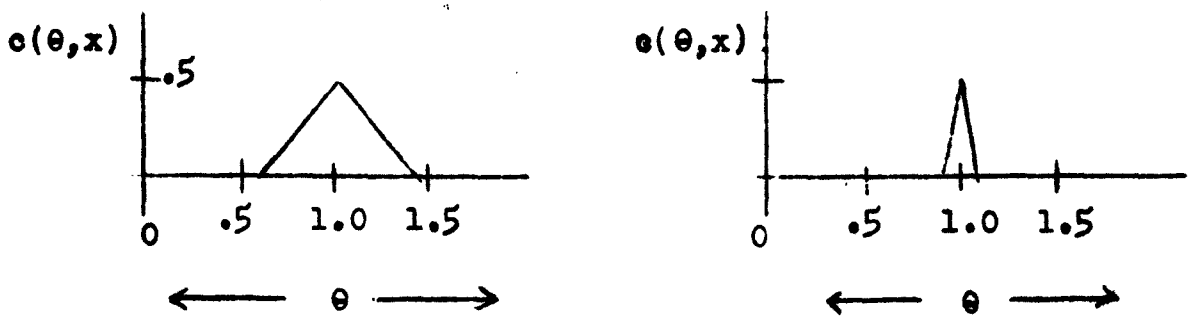
Alternatively, for any fixed k, $-\frac{1}{2} \leq k \leq \frac{1}{2}$, there is for each $H(\theta_0)$ and $H(\theta_0-)$ an admissible acceptance region

$$A(\theta_0) = \left\{ x \mid (\tfrac{1}{2} - k)r + (\tfrac{1}{2} + k)\, s \leqq \theta_0 + k \right\} .$$

From such tests we obtain admissible confidence limit estimators at each confidence level, and the corresponding admissible confidence curve estimator:

$$c(\theta,x) = \begin{cases} 0, \text{ if } \theta \geq r + \frac{1}{2} \text{ or } \theta \leq s - \frac{1}{2} , \\ \frac{1}{2}\left[1 - \frac{|r+s-2\theta|}{1-(s-r)}\right], \text{ otherwise} . \end{cases}$$

For the two samples considered above, we obtain the respective confidence curve estimates :



Since the last curve lies under that given by the first estimator for the same sample, it provides stronger inferences about θ. This is not inconsistent with the admissibility of the first estimator, which provides (at most confidence levels) stronger inferences (shorter confidence intervals) from relatively uninformative samples like the first sample.

Example 8. Cauchy median. Let Y have the Cauchy density function $h(y,\theta) = \frac{1}{\pi(1+(y-\theta)^2)}$, $-\infty < y < \infty$, $-\infty < \theta < \infty$. Then $S(y,\theta) = \frac{2(y-\theta)}{1+(y-\theta)^2}$. Taking $v(x,\theta) = S(y,\theta)$, the conditions of Corollary 1 are satisfied, and $v(x,\theta) = 0$ defines the median-unbiased locally-best estimator $\theta^*(y) = y$. However for $\alpha \neq \frac{1}{2}$, $0 < \alpha < 1$, the conditions of Corollary 1 are not satisfied by $v(x,\theta) = S(y,\theta) - G(\alpha)$. For $x = (y_1,y_2)$, even for $\alpha = \frac{1}{2}$, $v(x,\theta) = S(x,\theta) = \sum_{i=1}^{2} S(y_i,\theta)$ fails to satisfy the conditions of Corollary 1. (For $|y_2 - y_1|$ large, $S(x,\theta) = 0$ has three roots θ.) Thus in general there do not exist confidence limit estimators (nor median-unbiased estimators) which are locally-best uniformly in θ.

10. Introduction to general theory of admissible estimators.

To illustrate the general theory of admissible estimators, and the place of the methods introduced above within the general theory, we consider the case in which Ω is finite: $\Omega = \{\theta | \theta = 1,2,\ldots k\}$. The principal features of the general case (in which Ω is any subset of the real line) can be illustrated conveniently in this case, for which the complete theory can be developed by relatively elementary methods. For any such estimation problem, we have a specified family of density functions $f(x,\theta)$, $\theta = 1,\ldots,k$. For each estimator $\theta^*(x)$, let

$$b(u,\theta,\theta^*) = \begin{cases} \text{Prob}\,[\theta^*(X) = u|\theta], & \text{if } u \neq \theta \\ 0, & \text{if } u = \theta \end{cases},$$

for $u,\theta = 1,\ldots k$. The risk curves of θ^* are

$$a(u,\theta,\theta^*) = \begin{cases} \sum_{j \neq u} b(j,\theta,\theta^*), & \text{if } u < \theta, \\ 0, & \text{if } u = \theta, \\ \sum_{j \neq u} b(j,\theta,\theta^*), & \text{if } u > \theta \ . \end{cases}$$

It is useful to interpret such an estimation problem in relation to a somewhat different statistical inference or decision problem, which for brevity we shall call the multidecision problem: This other problem is that of choosing, on the basis of an observed value x, one of k specified simple hypotheses; it may also be described as an estimation problem which lacks a parametric structure in the sense that no ordering of the labels θ of the k hypotheses is relevant to the problem. Any measurable function $\theta^*(x)$ taking only the values $1,\ldots k$, represents both a possible solution to the multidecision problem (a decision function, or an inference function, or an "estimator" in the last-mentioned sense) and an estimator in the sense discussed above.

For the multidecision problem, the merits of each decision function $\theta^*(x)$ are represented completely by its error-probabilities $b(j,\theta,\theta^*)$; for each θ, such probabilities are the components of the vector-valued risk function of θ^* at θ. The general goal is to determine decision functions θ^* for which these error-probabilities are minimized jointly in some suitable sense. A decision function θ^* is called admissible if there is no other for which all corresponding error-probabilities are at least as small, with at least

one strictly smaller. Complete classes, minimal essentially complete classes, etc., are defined correspondingly (cf. Lindley [26] and Wolfowitz [27].)

A simple necessary condition that $\Theta^*(x)$ be admissible for the estimation problem is that it be admissible for the multidecision problem. For if Θ^{**} is better than Θ^* for the latter problem, $b(j,\theta,\Theta^{**}) \leqq b(j,\theta,\Theta^*)$ for all (j,θ), with at least one inequality strict; therefore $a(u,\theta,\Theta^{**}) \leqq a(u,\theta,\Theta^*)$ for all (u,θ), with at least one inequality strict. Thus the admissible estimators are a subclass (typically a relatively small one) of the admissible multidecision functions. Similarly every essentially complete class of multidecision functions contains an essentially complete class of estimators.

The relations between the estimation and multidecision problems can be illustrated further in terms of techniques, related to Bayes' formula, which play basic roles in the theory of each problem: For any estimation problem specified as above, let $q = q(u,\theta)$ be an arbitrary real-valued function such that $q(u,\theta) \geqq 0$ for $u, \theta = 1,\ldots k$; any such function will be called a <u>weight function</u> (for the estimation problem). For any such q and any estimator Θ^*, we define the (generalized) <u>Bayes risk</u>:

$$R(q,\Theta^*) = \sum_{\theta=1}^{k} \sum_{u=1}^{k} q(u,\theta)\, a(u,\theta,\Theta^*) \quad .$$

On the other hand, for any multidecision problem specified as above, let $Q = Q(u,\theta) \geqq 0$ be an arbitrary weight-function; then for any multidecision function Θ^* the corresponding Bayes risk is:

$$R'(Q,\theta^*) = \sum_{\theta=1}^{k} \sum_{u=1}^{k} Q(u,\theta)\, b(u,\theta,\theta^*) \ .$$

For any given θ^* and $q(u,\theta)$, we have

$$R(q,\theta^*) = \sum_{\theta} \left[\sum_{u>\theta} q(u,\theta) \sum_{j\geq u} b(j,\theta,\theta^*) + \sum_{u<\theta} q(u,\theta) \sum_{j\leq u} b(j,\theta,\theta^*) \right]$$

$$= \sum_{\theta} \left[\sum_{j>\theta} b(j,\theta,\theta^*) \sum_{u>\theta} q(u,\theta) + \sum_{j<\theta} b(j,\theta,\theta^*) \sum_{u<\theta} q(u,\theta) \right]$$

$$= \sum_{\theta} \sum_{j} Q(j,\theta),\ b(j,\theta,\theta^*) \ ,$$

where

$$Q(j,\theta) = \begin{cases} \sum_{j\geq u>\theta} q(u,\theta), & \text{for } j > \theta \ , \\ 0 \ , & \text{for } j = \theta \ , \\ \sum_{j\leq u<\theta} q(u,\theta), & \text{for } j < \theta \ . \end{cases}$$

For each θ, $Q(j,\theta)$ is nondecreasing in j for $j \geq \theta$, and nonincreasing in j for $j \leq \theta$; that is, $Q(j,\theta)$ has a single relative minimum which it assumes on one or more consecutive values of j including $j = \theta$. Thus each weight-function $q(u,\theta)$ for the estimation problem determines uniquely a weight-function $Q(j,\theta)$ for the multidecision problem, which has, for each θ, a single relative minimum. Conversely a weight-function $Q(j,\theta)$ for the multidecision problem having, for each θ, a single relative minimum (in the preceding sense) determines uniquely (through the last equation) a unique weight-function $q(u,\theta)$ for the estimation problem. Thus the Bayes solutions θ^* for the estimation problem (i.e. the functions

θ^* which, for some given q, minimize $R(q,\theta^*)$ are a subclass of the Bayes solutions for the multidecision problems, characterized by the preceding restriction on the possible forms of the weight function $Q(u,\theta)$ for the latter problem.

For any given weight-function q, the determination of Bayes estimators is conveniently carried out as follows: Let Q be determined by q as above. Then $R(q,\theta^*) = R'(Q,\theta^*)$ is minimized if, for each x, $\theta^*(x)$ takes the (a) value u which minimizes $\sum_{\theta=1}^{k} Q(u,\theta)\, f(x,\theta)$. A simple sufficient condition for admissibility of an estimator is that it be an essentially unique Bayes solution in the sense that for some q it minimizes $R(q,\theta^*)$, and every other estimator which also minimizes $R(q,\theta^*)$ has the same risk-curves $a(u,\theta)$. (A related sufficient condition for admissibility is that an estimator be a Bayes solution with respect to each of the weight functions $q_1,\ldots q_{r-1}$, and that among all <u>such</u> estimators it is an essentially unique Bayes solution with respect to some q_r.) Another simple sufficient condition for admissibility is that an estimator be a Bayes solution with respect to some q which is positive for all u,θ. Every admissible estimator is a Bayes solution with respect to some q; and the class of Bayes solution with respect to weight-functions q is a complete class of estimators.

Various specific formulations of the estimation problem can be exhibited as special cases of the present formulation. For example let $W(j,\theta)$ denote the loss function adopted in any decision-theoretic formulation: the loss incurred, if θ is true and it is inferred that $\theta = j$, is equal to $W(j,\theta)$. Then use of any estimator θ^* leads, when θ is true, to the expected loss

$$E[W(\theta^*(X),\theta)|\theta] = \sum_J b(j,\theta,\theta^*)\, W(j,\theta) = r(\theta,\theta^*) ,$$

a real-valued risk function (of θ). To illustrate the frequently adopted specification that losses are proportional to the <u>squared error</u> of the estimate, we replace the convenient labels $\theta = 1,\ldots k$ by the more general parameter values $\theta = \theta_1, \theta_2,\ldots\theta_k$, where $\theta_i < \theta_{i+1}$, and write $W(u,\theta_i) = c(\theta_i)\,(u - \theta_i)^2$, where u is any value in the range of θ^*. (The expected mean-squared error can generally be reduced further by dropping the restriction that the range of θ^* be the range of θ_i; the conflict between these considerations disappears in typical problems where the range of θ is an interval.) For any a priori probabilities g_i = Prob (θ_i), $i = 1,\ldots k$, any estimator θ^* gives the Bayes risk

$$\sum_{i=1}^{k} g_i r(\theta_i,\theta^*) = \sum_i g_i c(\theta_i) \sum_u b(u,\theta_i,\theta^*)(u - \theta_i)^2$$

$$= R'(Q,\theta^*) = R(q,\theta^*) ,$$

where $Q = Q(u,\theta_i) = g_i c(\theta_i)(u - \theta_i)^2$; $q(u,\theta_i)$ is determined by Q as above. Numerous examples are treated (without restrictions on $\cap$) in the texts and research literature of decision theory.

A <u>simple</u> loss function for the estimation problem is one of the form

$$W(j,\theta) = \begin{cases} 0, & \text{if } \theta_1(\theta) < j < \theta_2(\theta) , \\ c_1(\theta), & \text{if } j \leq \theta_1(\theta) , \\ c_2(\theta) & \text{if } j \geq \theta_2(\theta) , \end{cases}$$

where

$$c_1(\theta) \geq 0,\ \theta_1(\theta) \leq \theta \leq \theta_2(\theta),\ \theta_1(\theta) < \theta_2(\theta) \qquad \text{for } \theta = 1,2,\ldots,k.$$

This gives the risk function

$$r(\theta,\theta^*) = c_1(\theta)a(\theta_1(\theta),\theta,\theta^*) + c_2(\theta)a(\theta_2(\theta),\theta,\theta^*) .$$

If a priori probabilities $g(\theta)$ are adopted, then the Bayes risk, with the use of θ^*, is

$$\sum_{\theta} g(\theta)[c_1(\theta)a(\theta_1(\theta),\theta,\theta^*) + c_2(\theta)a(\theta_2(\theta),\theta,\theta^*)]$$

$$= R'(Q,\theta^*) = R(q,\theta^*) ,$$

where

$$q = q(u,\theta) = \begin{cases} g(\theta)c_1(\theta), & \text{if } u = \theta_1(\theta), \\ g(\theta)c_2(\theta), & \text{if } u = \theta_2(\theta), \\ 0, & \text{otherwise,} \end{cases}$$

and $Q(j,\theta)$ is determined by q as above.

The methods of Sections 6-9 above can be characterized in the present terms as follows: Writing

$$R(q,\theta^*) = \sum_u \left[\sum_{\theta>u} q(u,\theta)a(u,\theta,\theta^*) + \sum_{\theta\leq u} q(u+1,\theta)a(u+1,\theta,\theta^*) \right] ,$$

for each u the summand can be interpreted as a linear combination, with coefficients $q \geq 0$, of the various probabilities of errors of Types I and II given by a test of the one-sided hypothesis $H(u)$: $\theta \leq u$, against $H'(u)$: $\theta > u$, where the test has the acceptance

region $A(u) = \{x | \theta^*(x) \leqq u\}$. In other words, each such term (with index u) is the Bayes risk in a certain one-sided testing problem; it is minimized by a suitable acceptance region A(u) (determined by a technique equivalent to the Neyman-Pearson lemma); such Bayes acceptance regions are admissible under mild conditions. If the estimation problem has a suitably simple structure, and if the weight-function q is a suitable one, then the acceptance regions A(u) will constitute a nondecreasing sequence in u; in such cases, the Bayes risk in the estimation problem can be minimized by minimizing simultaneously each of the mentioned terms with respective indices $u = 1,\ldots,k$. The Bayes estimator obtained in such cases is:

$$\theta^*(x) = \begin{cases} u, & \text{if } x \in A(u) - A(u-1), \quad \text{for } u = 2,\ldots,k \\ 1, & \text{if } x \in A(1) . \end{cases}$$

It is problems having this structure which are treated in Section 6-9 above (without the restriction that Ω be finite). The method of Section 8 is represented by the form assumed by $R(q,\theta^*)$ for the special case of a simple loss-function, defined as above; in such cases the minimization of a term of R with index u corresponds to use of the Neyman-Pearson lemma to determine a best acceptance region A(u) for testing between two simple hypotheses.

If Ω is not finite, after choosing any finite subset $\Omega^* \subset \Omega$ (more or less "representative" of Ω) we can apply the above simple computational methods to determine Bayes estimators of $\theta \in \Omega^*$. If for any q, the Bayes estimator θ^* of $\theta \in \Omega^*$ is determined

essentially uniquely on the sample space (up to sets having probability 0 for all $\theta \in \Omega$), then θ^* is an admissible estimator of $\theta \in \Omega$. In this way, elementary techniques can provide a number of admissible estimators illustrative of the variety to be found in the full admissible class.

11. An application of the general theory; estimators having prescribed precision in a specified region; sequential probability ratio estimators.

It is sometimes desired that an estimator have high precision in some interval in the parameter space, while in the remainder of the parameter space much lower precision would suffice. In general efficient achievement of such a specification requires use of an estimator based on a sequential sampling rule. One formulation and solution of such a problem is the following; for illustrative purposes, a concrete example is discussed.

Let $Y_1, Y_2, \ldots$ be independent Bernoulli trials, with Prob $(Y_i = 1) = \theta$, Prob $(Y_i = 0) = (1 - \theta)$. An estimator θ^* is required which will have high precision for θ near .5. This requirement may be formulated in part as follows: For $\theta = .4$ or .6, the probability is at least .95 that θ^* will be closer to the correct one of these two values; in terms of risk curves of estimators, we require essentially that $a(.5,.4,\theta^*) \leqq .05$ and $a(.5,.6,\theta^*) \leqq .05$. (Further interpretations of these requirements in relation to the general notion of precision will appear below.) To meet these requirements, consider any estimator θ^*, and consider the test of the one-sided hypothesis H: $\theta \leqq .5$ against H': $\theta > .5$ given by the acceptance region $\{x | \theta^*(x) \leqq .5\}$. (The description

of the sample space on which our estimators are defined remains to be specified.) The requirements to be met by θ^* imply that this test has error-probabilities not exceeding .05 when $\theta = .4$ and $\theta = .6$. If sequential sampling rules are allowed, it is known that the last condition is satisfied most efficiently, in terms of expected number of observations Y_i required under $\theta = .4$ and $\theta = .6$, by Wald's sequential probability ratio test [28]. (We discuss such tests ignoring "excess at termination"; in problems of the type being considered, this entails that some of the following equations represent close approximations; for certain problems, no such qualification is necessary.) The indicated sampling rule is: Observe $Y_1, Y_2, \ldots$, compute after each observation Y_m the sum $d_m = \sum_{i=1}^{m} y_i$ and $h = h(m,d_m) = 2d_m - m$, and terminate observation as soon as either $h = k = \log(19)/\log(3/2)$ or $h = -k$. The resulting sample space is $S = \{x | x = (y_1, \ldots, y_n),\ n = 1,2,\ldots;\ |h(m,d_m)| < k \text{ or } = k \text{ as } m < n \text{ or } m = n\}$. The conditions specified above are met (with minimum expected sample sizes under all values of θ) by use of this sampling rule and any definition of $\theta^*(x)$ which satisfies:

$$\theta^*(x) \leq .5 \text{ for } x \text{ such that } h = -k$$

$$\theta^*(x) > .5 \text{ for } x \text{ such that } h = k.$$

The definition of $\theta^*(x)$ can be completed so as to make it admissible and median-unbiased. (Because S is discrete, use of an auxiliary randomization variable is necessary to obtain exact median-unbiasedness; we omit such randomization, obtaining an

admissible estimator which is approximately median-unbiased.) Every estimator satisfying the preceding inequalities is a Bayes solution for the above stated problem, given the sample space S. The determination of an admissible estimator among these can be interpreted as an illustration of the technique of using a sequence of a priori distributions; and of choosing, among all Bayes solutions for the first such distribution, one which minimizes the Bayes risk for the second such distribution.

We have

$$S^*(x,\theta) = d_n/\theta - (n-d_n)/(1-\theta)$$

$$= d_n/\theta(1-\theta) - n/(1-\theta)$$

$$= \begin{cases} n(\tfrac{1}{2} - \theta)/\theta(1-\theta) + k/2\theta(1-\theta), & \text{if } h = k, \\ n(\tfrac{1}{2} - \theta)/\theta(1-\theta) - k/2\theta(1-\theta), & \text{if } h = -k. \end{cases}$$

For any fixed $\theta_0 < \frac{1}{2}$, $S(x,\theta_0)$ is an increasing function of n as x varies subject to h = -k; and the set of such points has probability exceeding $\frac{1}{2}$ when $\theta = \theta_0$. To determine a test of $H(\theta_0)$: $\theta \leqq \theta_0$ against $H'(\theta_0)$: $\theta > \theta_0$, with acceptance region $\{x \mid \theta^*(x) \leqq \theta_0\}$, having size 1/2, and having the property that it is a locally-best test of this form subject to the conditions already imposed upon $\theta^*(x)$, it is necessary and sufficient that $\theta^*(x)$ satisfy the following additional condition: Let $n(\theta_0)$ be determined by

$$\text{Prob}\,(h = -k \text{ and } n \leqq n(\theta_0) \mid \theta_0) \doteq \tfrac{1}{2}.$$

In general, this relationship can be satisfied only approximately, but always closely except for θ_0 very near 0. Then

$$\theta^*(x) \leqq \theta_0 \text{ for } x \text{ such that } h = -k \text{ and } n \leqq n(\theta_0) \ ,$$

$$\theta^*(x) > \theta_0 \quad \text{otherwise.}$$

As θ_0 increases from 0 to $\frac{1}{2}$, $n(\theta_0)$ takes successively the values $k, k+1, k+2, \ldots$.

Proceeding similarly for any fixed $\theta_0 > \frac{1}{2}$, we define $n(\theta_0)$ similarly for such values, and obtain the conditions

$$\theta^*(x) \leqq \theta_0 \text{ for } x \text{ such that } h = k \text{ and } n \geqq n(\theta_0) \ ,$$

$$\theta^*(x) > \theta_0 \quad \text{otherwise} \ .$$

It is clear that all of these conditions on $\theta^*(x)$ can be met simultaneously (allowing the approximations mentioned), and that they provide a full definition of the estimator. Since this definition depends on x only through $n = n(x)$ and $h = h(x) = \pm k$, θ^* depends on x only through $t = t(x) = h/kn$. The range of t is $\pm 1, \pm 1/2, \pm 1/3, \ldots$ and θ^* is an increasing function of t.

Let $F(t,\theta) = \text{Prob}\left\{t(X) \leqq t \middle| \theta\right\}$, then the estimator $\theta^* = \theta^*(x,.5)$ is defined as the root θ of the equation

$$v(x,\theta,.5) \equiv F(t(x),\theta) - .5 = 0 \ .$$

More generally, for each α, $0 < \alpha < 1$, a confidence limit estimator $\theta^*(x,\alpha)$ is defined as the root θ of $v(x,\theta,\alpha) = 0$. (The admissibility of such estimators can be shown as above.) The family of such estimators constitutes an admissible confidence curve estimator.

Confidence curve estimates of this kind will be narrow, reflecting high precision, when n is very large, and will be wide reflecting low precision, when n is very small. It follows from the requirements imposed upon $\theta^*(x,.5)$ above that whenever $\theta^*(x,.5) > .5$, we have $\theta^*(x,.95) > .4$ (whether n is small or large), and that whenever $\theta^*(x,.5) < .5$, we have $\theta^*(x,.05) < .6$; hence the 90 percent confidence interval $J(x) = [\theta^*(x,.95), \theta^*(x,.05)]$ will never include both the values $\theta = .4$ and $\theta = .6$. (The event $n(x) = +\infty$, which has probability 0 under each θ, gives $J(x) = [.4,.6]$ and $\theta^*(x,.5) = .5$.) This constitutes a useful interpretation of the formulation adopted above of the general requirement of high precision for θ near .5.

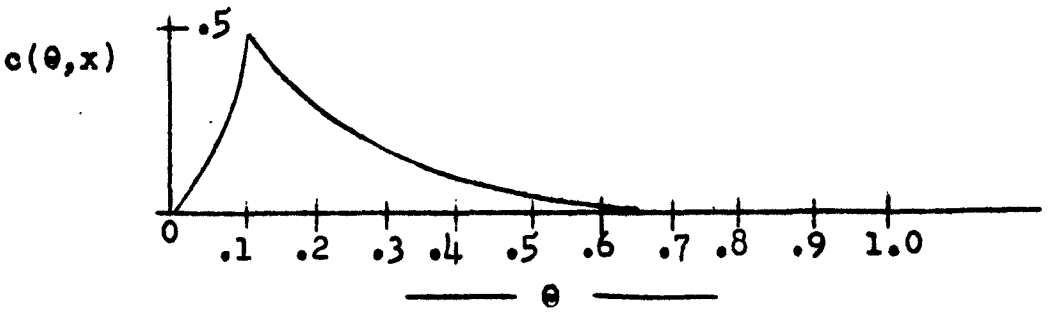
For practical reasons, it is sometimes necessary to terminate sampling before this is indicated by the above sampling rule, and the question arises what inferences can be made validly on the basis of such partial determination of an observation x. Termination after m observations with $|h(m,d_m)| < k$ is equivalent to observation of the event $-1/m < t(x) < 1/m$. For each α, this implies that the estimate $\theta^*(x,\alpha)$ (which would have been determined by continuing sampling) satisfies $\underline{\theta}^*(x,\alpha) < \theta^*(x,\alpha) < \overline{\theta}^*(x,\alpha)$, where $\underline{\theta}^*(x,\alpha)$ are respectively the roots θ of $F(-1/m,\theta) = \alpha$ and

of $F(1/m,\theta) = \alpha$. These bounds on an estimate narrow progressively with increasing m. When such bounds on an estimate (or confidence curve) become sufficiently narrow for the purpose at hand, sampling can be terminated without affecting the validity of the (approximate) estimates obtained.

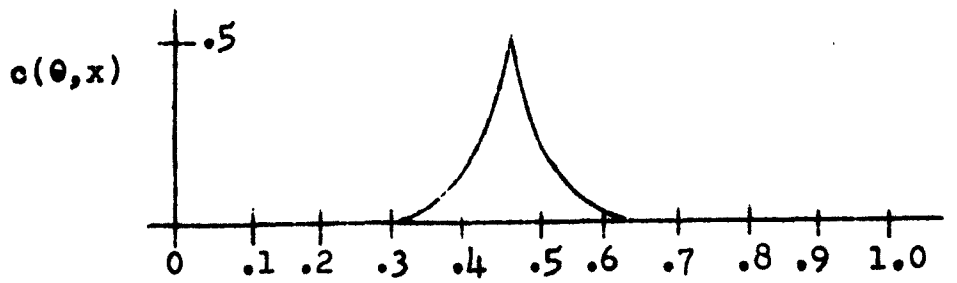
Concerning the computation of values of $F(t,\theta)$ required for use of such estimators, the function $F(0,\theta)$ of θ is the operating characteristic function of a sequential probability ratio test, on which there is an extensive theoretical and quantitative literature for a wide range of problems. For each θ, when $F(0,\theta)$ is known, the determination of $F(t,\theta)$ is reduced to the problem of determining the conditional cumulative distribution function of n (the number of observations required for termination of sampling, or the duration of a random walk with two absorbing barriers) on the condition of termination with $h = -k$ ("acceptance of H: $\theta \leq .5$", or absorption at the lower boundary), and again on the condition of termination with $h = k$ ("rejection of H", or absorption at the upper boundary). (The unconditional distribution of n, together with one of these conditional distributions and $F(0,\theta)$, determines the other conditional distribution.)

SCHEMATIC ILLUSTRATIONS OF CONFIDENCE CURVE ESTIMATES OF A BINOMIAL PARAMETER θ HAVING HIGH PRECISION FOR θ NEAR $\frac{1}{2}$

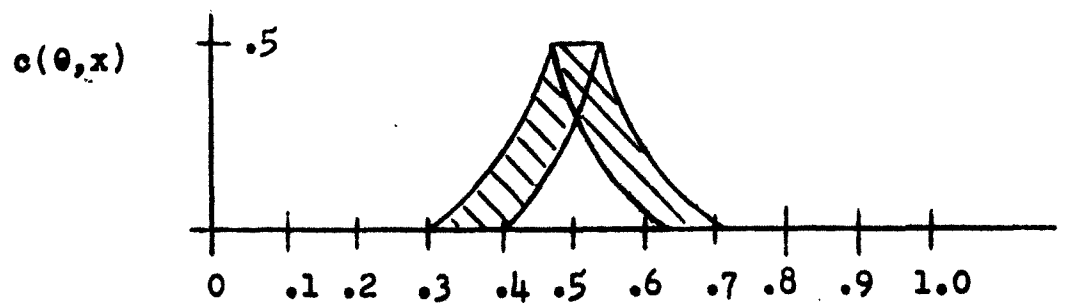
(A) n(x) very small, h(x) = -k



(B) n(x) very large, h(x) = -k



(C) Bounds on estimate; sampling curtailed with m very large.



REFERENCES

[1] Lindley, D. V. "Statistical Inference." Journal of the Royal Statistical Society. Series B, v. XV, 1953, pp. 30-76.

[2] Neyman, J. "Current Problems of Mathematical Statistics." Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, v. I, pp. 1-22.

[3] Walsh, C. M. The Problem of Estimation. A Seventeenth-century Controversy and its Bearings on Modern Statistical Questions, Especially Index-numbers, P. S. King and Son, Ltd., London, 1921.

[4] Savage, L. J. Foundations of Statistics, John Wiley and Sons, Inc., New York 1954.

[5] Pitman, E. J. G. "The estimation of the location and scale parameters of a continuous population of any form." Biometrika, v. 30, 1939, pp. 391-421.

[6] Weiss, L. "A Higher Order Complete Class Theorem." Annals of Math. Stat., v. 24, 1953, pp. 677-680.

[7] Lehmann, E. Testing Statistical Hypotheses, J. Wiley, 1959.

[8] Neyman, J. "Outline of a theory of statistical estimation based on the classical theory of probability," Philos. Trans. of the Royal Society of London, Ser. A, No. 767, v. 236, 1937, pp. 333-360.

[9] Wolfowitz, J. "Minimax estimates of the mean of a normal distribution with known variance." Annals of Math. Stat., v. 21, 1950, pp. 218-230.

[10] Tukey, John W. "Standard confidence points." Memorandum Report 26, Statistical Research Group, Princeton University, July 26, 1949. (Preliminary report presented at meeting of the Institute of Mathematical Statistics, New York City, April 1948.)

[11] Cox, D. R. "Some Problems Connected with Statistical Inference." Annals of Math. Stat., v. 29, 1958, pp. 357-372.

[12] Lehmann, E. Testing Statistical Hypotheses, J. Wiley, 1959.

[13] Cramer, H. Mathematical Methods of Statistics, Princeton, 1946.

[14] Wald, A. "Asymptotically shortest confidence intervals," Ann. Math. Stat., v. 13, 1942, pp. 127-137.

[15] Lehmann, E. "Some comments on large sample tests," Proc. of the Berkeley Symposium on Mathematical Statistics and Probability, U. of California Press, 1949, pp. 451-458.

[16] Eisenhart, Churchill, and Martin, Celia S. "The relative frequencies with which certain estimators of the standard deviation of a normal population tend to underestimate its value."
Abstract and tables distributed at meeting of the Institute of Mathematical Statistics, Madison, Wisconsin, Sept. 7, 1948. Abstract in Annals of Mathematical Statistics, v. 19, 1948, p. 600.

[17] Berkson, J. "Tables for the maximum likelihood estimate of the logistic function." Biometrika, v. 13, 1957, pp. 28-34.

[18] Wallace, David. "Asymptotic approximations to distributions," Annals of Math. Stat., v. 29, 1958, pp. 635-654.

[19] Cochran, W. C. "Estimation of bacterial densities by means of 'the most probable number'." Biometrics, v. 6, 1950, pp. 105-116.

[20 Lord, F. "An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability." Psychometrika, v. 18, 1953, pp. 57-76.

[21] Birnbaum, A. "Statistical theory of tests of a mental ability." To be published. Abstract in Annals of Math Stat., v. 29, 1958, p. 1285.

[22] Finney, D. J. Probit Analysis. Cambridge, 1952.

[23] Berkson, Joseph. "A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function." Journal of the American Statistical Association, v. 48, 1953, pp. 565-599.

[24] Pratt, J. W. "Admissible one-sided tests for the mean of rectangular distribution," "Annals of Math. Stat." v. 29, 1958, pp. 1268-1271.

[25] Venkataraman, Lakshmi, Note on complete classes of admissible estimators of the mean of a rectangular distribution. To be published.

[26] Wolfowitz, J. Review of [1] by D. V. Lindley. Math. Reviews, v. 15, 1954, p. 242.

[27] Wald, A., and Wolfowitz, J. "Optimum character of the sequential probability ratio test." Annals of Math. Stat., v. 19, 1948, pp. 326-339.